What are the Differences between Bayesian Classifiers and Mutual-Information Classifiers?

Bao-Gang Hu, *Senior Member, IEEE*

***Abstract*—In this study, both Bayesian classifiers and mutual-information classifiers are examined for binary classifications with or without a reject option. The general decision rules in terms of distinctions on error types and reject types are derived for Bayesian classifiers. A formal analysis is conducted to reveal the parameter redundancy of cost terms when abstaining classifications are enforced. The redundancy implies an intrinsic problem of "*non-consistency*" for interpreting cost terms. If no data is given to the cost terms, we demonstrate the weakness of Bayesian classifiers in class-imbalanced classifications. On the contrary, mutual-information classifiers are able to provide an objective solution from the given data, which shows a reasonable balance among error types and reject types. Numerical examples of using two types of classifiers are given for confirming the theoretical differences, including the extremely-class-imbalanced cases. Finally, we briefly summarize the Bayesian classifiers and mutual-information classifiers in terms of their application advantages, respectively.**

***Index Terms*—Bayes, entropy, mutual information, error types, reject types, abstaining classifier, cost sensitive learning.**

I. Introduction

The Bayesian principle provides a powerful and formal means of dealing with statistical inference in data processing, such as classifications [1]. If classifiers are designed based on this principle, they are called "*Bayesian classifiers*" in this work. The learning targets for Bayesian classifiers are either the minimum error or the lowest cost. It was recognized that Chow [2][3] was "*among the earliest to use Bayesian decision theory for pattern recognition*" [4]. His pioneering work is so enlightening that its idea of optimal tradeoff between error and reject still sheds a bright light for us to deep our understanding to the subject, as well as to explore its applications widely in this information-explosion era. In recent years, cost sensitive learning and class-imbalanced learning have received much attentions in various applications [12-18]. For classifications of imbalanced, or skewed, datasets, "*the ratio of the small to the large classes can be drastic such as 1 to 100, 1 to 1,000, or 1 to 10,000 (and sometimes even more)*" [16]. It was pointed out by Yang and Wu [19] that dealing with imbalanced and cost-sensitive data is among the ten most challenging problems in the study of data mining. In fact, the related subjects are not a new challenge but a more crucial concern than before for increasing needs of searching useful information from massive data. Binary classifications will be a basic problem in such application background. Classifications based on cost terms for the tradeoff of error types is a conventional subject in medical diagnosis. Misclassification from "*type I error*" (or "*false positive*") or from "*type II error*" (or "*false negative*") is significantly different in the context of medical practices. In other domains of applications, one also needs to discern error types for attaining reasonable results in classifications. Among all these investigations, cost terms, which is usually specified by users from a cost matrix, play a key role in class-imbalanced learning [11-14][20][46][47].

In binary classifications with a reject option, Bayesian classifiers require a cost matrix with six cost terms as the given data. Different from the prior to the probabilities of classes, this requirement can be another source of subjectivity that disqualifies Bayesian classifiers as an objective approach of induction [43]. If an objectivity aspect is enforced for classifications with a reject option, a difficulty does exist for Bayesian classifiers that assign cost terms objectively. The cost terms for error types may be given from an application background, but are generally unknown for reject types. In binary classifications, Chow [3] and early researchers [22][23][24] usually assumed no distinctions among errors and among rejects. The later study in [31] considered different costs for correct classification and miscalssifications, but not for rejects. The more general settings for distinguishing error types and reject types were reported in [25][27][28]. To overcome the problems of presetting cost terms manually, Pietraszek [28] proposed two learning models, namely, "*bounded-abstention*" and "*bounded-improvement*", and Grall-Maës and Beauseroy [30] applied a strategy of adding performance constraints for class-selective rejection. If constraints either on total reject or on total error, they may result in no distinctions between their associated cost terms. Up to now, it seems that no study has been reported for the objective design of Bayesian classifiers by distinguishing error types and reject types at the same time.

Several investigations are reported by following Chow's rule on classifier designs with a reject option [21-30]. In addition to a kind of "*ambiguity reject*" studied by Chow, the other kind of "*distance reject*" was also considered in [21]. Ambiguity reject is made to a pattern located in an ambiguous region between/among classes. Distance reject represents a pattern far away from the means of any class and is conventionally called an "*outlier*" in statistics [4]. Ha [22] proposed another important kind of reject, called "*class-selective reject*", which defines a subset of classes. This scheme is more suitable to multiple-class classifications. For example, in three-class problems, Ha's classifiers will output the predictions including "*ambiguity reject between Class 1 and 2*", "*ambiguity reject among Class 1, 2 and 3*", and the other rejects from class

Bao-Gang Hu is with NLPR/LIAMA, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
E-mail: hubg@nlpr.ia.ac.cn

combinations. Multiple rejects with such distinctions will be more informative than a single "*ambiguity reject*". Among all these investigations, the Bayesian principle is applied again for their design guideline of classifiers.

While the Bayesian inference principle is widely applied in classifications, another principle based on the mutual information concept is rarely adopted for designing classifiers. Mutual information is one of the important definitions in entropy theory [38]. Entropy is considered as a measure of uncertainty within random variables, and mutual information describes the relative entropy between two random variables [9]. If classifiers seek to maximize the relative entropy for their learning target, we refer them to "*mutual-information classifiers*". It seems that Quinlan [5] was among the earliest to apply the concept of mutual information (but called "*information gain*" in his famous ID3 algorithm) in constructing the decision tree. Kvålseth [6] and Wickens [7] introduced the definition of normalized mutual information (NMI) for assessing a contingency table, which laid down the foundation on the relationship between a confusion matrix and mutual information. Being pioneers in using an information-based criterion for classifier evaluations, Kononenko and Bratko [41] suggested the term "*information score*" which was equivalent to the definition of mutual information. A research team leaded by Principe [8] proposed a general framework, called "*Information Theoretic Learning* (ITL)", for designing various learning machines, in which they suggested that mutual information, or other information theoretic criteria, can be set as an objective function in classifier learning. Mackay [[9], page 533] once showed numerical examples for several given confusion matrices, and he suggested to apply mutual information for ranking the classifier examples. Wang and Hu [10] derived the nonlinear relations between mutual information and the conventional performance measures, such as accuracy, precision, recall and F1 measure for binary classifications. In [11], a general formula for normalized mutual information was established with respect to the confusion matrix for multiple-class classifications with/without a reject option, and the advantages and limitations of mutual-information classifiers were discussed. However, no systematic investigation is reported for a theoretical comparison between Bayesian classifiers and mutual-information classifiers in the literature.

This work focuses on exploring the theoretical differences between Bayesian classifiers and mutual-information classifiers in classifications for the settings with/without a reject option. In particular, this paper derives much from and consequently extends to Chow's work by distinguishing error types and reject types. To achieve analytical tractability without losing the generality, a strategy of adopting the simplest yet most meaningful assumptions to classification problems is pursued for investigations. The following assumptions are given in the same way as those in the closed-form studies of Bayesian classifiers by Chow [3] and Duda, et al [4]:

A1. Classifications are made for two categories (or classes) over the feature variables.
A2. All probability distributions of feature variables are exactly known.

One may argue that the assumptions above are extremely restricted to offer practical generality in solving real-world problems. In fact, the power of Bayesian classifiers does not stay within their exact solutions to the theoretical problems, but appear from their generic inference principle in guiding real applications, even in the extreme approximations to the theory. We fully recognize that the assumption of complete knowledge on the relevant probability distributions may be never the cases in real-world problems [31][33]. The closed-form solutions of Bayesian classifiers on binary classifications in [3][4] have demonstrated the useful design guidelines that are applicable to multiple classes [22]. The author believes that the analysis based on the assumptions above will provide sufficient information for revealing the theoretical differences between Bayesian classifiers and mutual-information classifiers, while the intended simplifications will benefit readers to reach a better, or deeper, understanding to the advantages and limitations of each type of classifiers.

The contributions of this work are twofold. First, the analytical formulas for Bayesian classifiers and mutual-information classifiers are derived to include the general cases with distinctions among error types and reject types for cost sensitive learning in classifications. Second, comparisons are conducted between the two types of classifiers for revealing their similarities and differences. Specific efforts are made on a formal analysis of parameter redundancy to the cost terms for Bayesian classifiers when a reject option is applied. Section II presents a general decision rule of Bayesian classifiers with or without a reject option. Sections III provides the basic formulas for mutual-information classifiers. Section IV investigates the similarities and differences between two types of classifiers, and numerical examples are given to highlight the distinct features in their applications. The question presented in the title of the paper is concluded by a simple answer in Section V.

II. Bayesian Classifiers with a Reject Option

A. General Decision Rule for Bayesian Classifiers

Let $\mathbf{x}$ be a random pattern satisfying $\mathbf{x} \in \mathbf{X} \subset R^d$, which is in a d-dimensional feature space and will be classified. The true (or target) state t of $\mathbf{x}$ is within the finite set of two classes, $t \in T = \{t_1, t_2\}$, and the possible decision output $y = f(\mathbf{x})$ is within three classes, $y \in Y = \{y_1, y_2, y_3\}$, where f is a function for classifications and y_3 represents a "*reject*" class. Let $p(t_i)$ be the prior probability of class t_i and $p(\mathbf{x}|t_i)$ be the conditional probability density function of $\mathbf{x}$ given that it belongs to class t_i. The *posterior* probability $p(t_i|\mathbf{x})$ is calculated through the Bayes formula [4]:

$$p(t_i|\mathbf{x}) = \frac{p(\mathbf{x}|t_i)p(t_i)}{p(\mathbf{x})}, \qquad (1)$$

where $p(\mathbf{x})$ represents the mixture density for normalizing the probability. Based on the posterior probability, the Bayesian rule assigns a pattern $\mathbf{x}$ into the class that has the highest posterior probability. Chow [2][3] first introduced the framework of the Bayesian decision theory into the study of pattern recognition and derived the best error-type trade-off formulas

and the related optimal reject rule. The purpose of the reject rule is to minimize the total risk (or cost) in classifications. Suppose λ_{ij} is a cost term for the true class of a pattern to be t_i, but decided as y_j. Then, the conditional risk for classifying a particular $\mathbf{x}$ into y_j is defined as:

$$Risk(y_j|\mathbf{x}) = \sum_{i=1}^{2} \lambda_{ij} p(t_i|\mathbf{x}) = \sum_{i=1}^{2} \lambda_{ij} \frac{p(\mathbf{x}|t_i)p(t_i)}{p(\mathbf{x})}, \quad j = 1, 2, 3. \tag{2}$$

Note that the definition of λ_{ij} in this work is a bit different with that in [4], so that λ_{ij} will form a 2×3 matrix. Chow [3] assumed the initial constraints on λ_{ij} from the intuition in classifications:

$$\lambda_{ik} > \lambda_{i3} > \lambda_{ii} \geq 0, \quad i \neq k, \quad i = 1, 2, \quad k = 1, 2. \tag{3}$$

The constraints imply that a misclassification will suffer a higher cost than a rejection, and a rejection will cost more than a correct classification. Relations about λ_{ij} are the main concern in the study of cost-sensitive learning, and this issue will be addressed later in this work. The total risk for the decision output y will be [4]:

$$Risk(y) = \int_V \sum_{j=1}^{3} \sum_{i=1}^{2} \lambda_{ij} p(t_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}, \tag{4}$$

with integration over the entire observation space V.

Definition 1 (Bayesian classifier): If a classifier is determined from the minimization of its risk over all patterns:

$$y^* = arg \min_y Risk(y), \tag{5a}$$

or in anther form on a given pattern $\mathbf{x}$:

$$Decide \quad y_j \quad if \quad Risk(y_j|\mathbf{x}) = \min_i Risk(y_i|\mathbf{x}) \tag{5b}$$

this classifier is called "*Bayesian classifier*", or "*Chow's abstaining classifier*" [27]. The term of $Risk(y^*)$ is usually called "*Bayesian risk*", or "*Bayesian error*" in the cases that zero-one cost terms ($\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = \lambda_{21} = 1$) are used for no rejection classifications[4].

In [3], a single threshold for a reject option was investigated. This setting was obtained for the assumption that cost terms are applied without distinction among the errors and among rejects. Following Chow's approach but with extension to the general cases to cost terms, one is able to derive the general decision rule on the rejection for Bayesian classifiers.

Theorem 1: The general decision rule for Bayesian classifiers are:

$$\begin{aligned} &Decide\ y_1\ if\ \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} > \delta_1, \\ &No\ rejection:\ \delta_1 = \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}}, \\ &Rejection:\ \delta_1 = \frac{\lambda_{21} - \lambda_{23}}{\lambda_{13} - \lambda_{11}}, \end{aligned} \tag{6a}$$

$$\begin{aligned} &Decide\ y_2\ if\ \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} \leq \delta_2, \\ &No\ rejection:\ \delta_2 = \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}}, \\ &Rejection:\ \delta_2 = \frac{\lambda_{23} - \lambda_{22}}{\lambda_{12} - \lambda_{13}}, \end{aligned} \tag{6b}$$

$$\begin{aligned} &Decide\ y_3\ if\ \frac{T_{r2}}{1 - T_{r2}} = \frac{\lambda_{23} - \lambda_{22}}{\lambda_{12} - \lambda_{13}} \\ &< \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} \leq \frac{\lambda_{21} - \lambda_{23}}{\lambda_{13} - \lambda_{11}} = \frac{1 - T_{r1}}{T_{r1}}, \end{aligned} \tag{6c}$$

$$\begin{aligned} &Subject\ to\ 0 < \frac{\lambda_{23} - \lambda_{22}}{\lambda_{12} - \lambda_{13}} < \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}} \\ &< \frac{\lambda_{21} - \lambda_{23}}{\lambda_{13} - \lambda_{11}}, \quad and \end{aligned} \tag{6d}$$

$$\begin{aligned} &No\ rejection:\ T_{r1} = T_{r2} = 0.5, \\ &Rejection:\ 0 < T_{r1} + T_{r2} \leq 1. \end{aligned} \tag{6e}$$

Eq (6c) applies the definition of two thresholds (called "*rejection thresholds*" in [3]), T_{r1} and T_{r2}.

Proof: See Appendix A. ■

Note that eq. (6d) suggests general constraints over λ_{ij}. The necessity for having such constraints is explained in Appendix A. A graphical interpretation to the two thresholds is illustrated in Fig. 1. Based on eq. (6c), the thresholds can be calculated from the following formulas:

$$\begin{aligned} T_{r1} &= \frac{\lambda_{13} - \lambda_{11}}{\lambda_{13} - \lambda_{11} + \lambda_{21} - \lambda_{23}},\ and \\ T_{r2} &= \frac{\lambda_{23} - \lambda_{22}}{\lambda_{12} - \lambda_{13} + \lambda_{23} - \lambda_{22}}. \end{aligned} \tag{7}$$

Eq. (7) describes general relations between thresholds and cost terms on binary classifications, which enables the classifiers to make the distinctions among errors and among rejects. Note that the special settings of Chow's rules [3] can be derived from eq. (7):

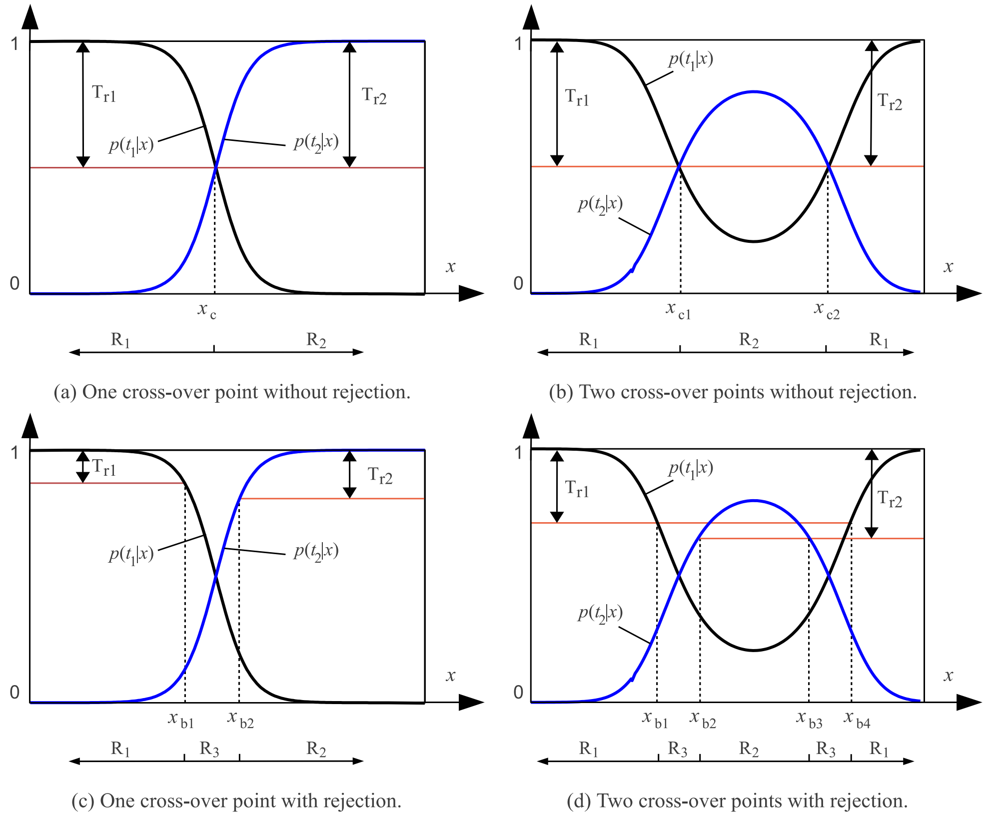
$$\lambda_{11} = \lambda_{22} = 0,\ \lambda_{12} = \lambda_{21} = 1,\ \lambda_{13} = \lambda_{23} = T_r. \tag{8}$$

Another important relation in [28] can also be obtained:

$$\begin{aligned} &\lambda_{11} = \lambda_{22} = 0, \\ &0 < \lambda_r = \lambda_{13} = \lambda_{23} < \frac{\lambda_{12}\lambda_{21}}{\lambda_{12} + \lambda_{21}}, \\ &T_{r1} = \frac{\lambda_r}{\lambda_{21}} \quad and \quad T_{r2} = \frac{\lambda_r}{\lambda_{12}}. \end{aligned} \tag{9}$$

Pietraszek [28] derived the rational region of λ_r above through ROC curves. The error costs can be different but not for reject ones. Note that, however, the rejection thresholds will be different when $\lambda_{12} \neq \lambda_{21}$. For advanced applications, Vanderlooy, et al [29] generalized Chow's rules by distinguishing error types and reject types, and derived the relations between two "*likelihood ratio thresholds*" and cost terms. Their rules of missing the terms λ_{11} and λ_{22} are not theoretically general, yet sufficient for applications. They derived formulas only from the inequality constraints of $Risk(y_1|\mathbf{x}) > Risk(y_3|\mathbf{x})$ and $Risk(y_2|\mathbf{x}) > Risk(y_3|\mathbf{x})$, respectively. Up to now, it seems no one has reported the general constraints (6d) in the literature. Based on eq. (6d), one can derive the rational (3), rather than employing the intuition.

By applying eq. (1) and the constraint $p(t_1|\mathbf{x}) + p(t_2|\mathbf{x}) = 1$, one can achieve the decision rules from eq. (6) with respect



(a) One cross-over point without rejection.
(b) Two cross-over points without rejection.
(c) One cross-over point with rejection.
(d) Two cross-over points with rejection.

Fig. 1. Rejection scenarios from the plots of $p(t_i|x)$ for univariate Gaussian distributions.

to the posterior probabilities and thresholds in a simple and better form for abstaining classifiers:

$$\begin{array}{l} Decide\ y_1\ if\ p(t_1|\mathbf{x}) > 1 - T_{r1}, \\ Decide\ y_2\ if\ p(t_2|\mathbf{x}) \geq 1 - T_{r2}, \\ Decide\ y_3\ for\ otherwise, \\ \quad Subject\ to\ 0 < T_{r1} + T_{r2} \leq 1. \end{array} \tag{10}$$

In comparison with the decision rules of eq. (6), which are expressed in terms of the likelihood ratio, eq. (10) together with Fig. 1 presents a better view for users to understand abstaining Bayesian classifiers. A plot of posterior probabilities show advantages over a plot of the likelihood ratio (Figure 2.3 in [4]) for determining rejection thresholds. Note that in Fig. 1 the plots are depicted on a one-dimensional variable for Gaussian distributions of X. The simplification supports the suggestions by Duda, et al, that one "*should not obscure the central points illustrated in our simple example*" [4]. Two sets of geometric points are shown for the plots. One set is called "*cross-over points*", denoted by x_{ci}, which are formed from two curves of $p(t_1|x)$ and $p(t_2|x)$. And the other is termed "*boundary points*", denoted by x_{bj}. The boundary points partition classification regions for one-dimensional problems. For a "*no rejection*" case, the boundary points are controlled by the ratio of $(\lambda_{21} - \lambda_{22})/(\lambda_{12} - \lambda_{11})$. In abstaining classifications, those points are determined from two thresholds, respectively. For multiple dimension problems, one can understand that both types of the points above become to be curves or even hypersurfaces.

With the exact knowledge of $p(t_i)$, $p(\mathbf{x}|t_i)$, and λ_{ij}, one can calculate Bayesian risk from the following equation:

$$\begin{array}{l} Risk(y^*) = \lambda_{11} CR_1 + \lambda_{12} E_1 + \lambda_{13} Rej_1 + \lambda_{22} CR_2 \\ +\lambda_{21} E_2 + \lambda_{23} Rej_2 \\ = \lambda_{11} \int\limits_{R_1} p(t_1) p(\mathbf{x}|t_1) d\mathbf{x} + \lambda_{12} \int\limits_{R_2} p(t_1) p(\mathbf{x}|t_1) d\mathbf{x} \\ +\lambda_{13} \int\limits_{R_3} p(t_1) p(\mathbf{x}|t_1) d\mathbf{x} + \lambda_{21} \int\limits_{R_1} p(t_2) p(\mathbf{x}|t_2) d\mathbf{x} \\ +\lambda_{22} \int\limits_{R_2} p(t_2) p(\mathbf{x}|t_2) d\mathbf{x} + \lambda_{23} \int\limits_{R_3} p(t_2) p(\mathbf{x}|t_2) d\mathbf{x}, \end{array} \tag{11}$$

where CR_i, E_i and Rej_i are the probabilities of "*Correct Recognition*", "*Error*", and "*Rejection*" for the ith class in the classifications, respectively; and R_1 to R_3 are the classification regions of Class 1, Class 2 and the reject class, respectively. The general relations among CR_i, E_i and Rej_i for binary classifications are given by [3]:

$$\begin{array}{l} CR_1 + CR_2 + E_1 + E_2 + Rej_1 + Rej_2 \\ \qquad\qquad = CR + E + Rej = 1, \\ and \quad A = \dfrac{CR}{CR + E}, \end{array} \tag{12}$$

where CR, E, and Rej represent total correct recognition, total error and total reject rates, respectively; and A is the accuracy rate of classifications.

B. Parameter Redundancy Analysis of Cost Terms

Bayesian classifiers present one of the general tools for cost sensitive learning. From this perspective, there exists a need

for a systematic investigation into a parameter redundancy analysis of cost terms for Bayesian classifiers, which appears missing for a reject option. This section will attempt to develop a theoretical analysis of parameter redundancy for cost terms.

For Bayesian classifiers, when all cost terms are given along with the other relevant knowledge about classes, a unique set of solutions will be obtained. However, this phenomenon does not indicate that all cost terms will be independent for determining the final results of Bayesian classifiers. In the followings, a parameter dependency analysis is conducted because it suggests a theoretical basis for a better understanding of relations among the cost terms and the outputs of Bayesian classifiers. Based on [35][36], we present the relevant definitions but derive a theorem from the functionals in eqs. (4) and (5) so that it holds generality for any distributions of features. Let a parameter vector be defined as $\theta = \{\theta_1, \theta_2, \cdots, \theta_p\} \in \mathbf{S}$, where p is the total number of parameters in a model $f(\mathbf{x}, \theta)$ and $\mathbf{S}$ denotes the parameter space.

Definition 2 (Parameter redundancy [35]): A model $f(\mathbf{x}, \theta)$ is considered to be parameter redundant if it can be expressed in terms of a smaller sized parameter vector $\beta = \{\beta_1, \beta_2, \cdots, \beta_q\} \in \mathbf{S}$, where $q < p$.

Definition 3 (Independent parameters): A model $f(\mathbf{x}, \beta)$ is said to be governed by independent parameters if it can be expressed in terms of the smallest size of parameter vector $\beta = \{\beta_1, \beta_2, \cdots, \beta_m\} \in \mathbf{S}$. Let $N_{IP}(\beta)$ denote the total number $(= m)$ of β for the model $f(\mathbf{x}, \beta)$.

Definition 4: (Function of parameters, parameter composition, input parameters, intermediate parameters): Suppose three sets of parameter vectors are denoted by $\theta = \{\theta_1, \theta_2, \cdots, \theta_p\} \in \mathbf{S}_1$, $\gamma = \{\gamma_1, \gamma_2, \cdots, \gamma_q\} \in \mathbf{S}_2$, and $\eta = \{\eta_1, \eta_2, \cdots, \eta_r\} \in \mathbf{S}_3$. If for a model there exists $f(\mathbf{x}, \theta) = f(\mathbf{x}, \varphi(\psi(\theta)))$ for φ: $\mathbf{S}_1 \to \mathbf{S}_2$ and ψ: $\mathbf{S}_2 \to \mathbf{S}_3$, we call φ and ψ to be functions of parameters, and $\varphi(\psi(\theta))$ to be parameter composition, where θ_i are called input parameters for $f(\mathbf{x}, \varphi(\psi(\theta)))$, γ_j and η_k are intermediate parameters.

Lemma 1: Suppose a model holds the relation $f(\mathbf{x}, \theta) = f(\mathbf{x}, \varphi(\psi(\theta)))$ for Definition 4. The total number of independent parameters of θ, denoted as $N_{IP}(f, \theta)$ for the model f will be no more than $\min(p, q, r)$, or in a form of:

$$N_{IP}(f, \theta) \leq \min(p, q, r) \tag{13}$$

Proof: Suppose $f(\mathbf{x}, \theta = \{\theta_1, \theta_2, \cdots, \theta_p\})$ without parameter composition, one can prove that $N_{IP}(f, \theta) \leq \min(p)$. According to Definition 2, any increase of its size of θ over p will produce a parameter redundancy in the model. Definition 3 indicates that the vector size p will be an upper bound for $N_{IP}(f, \theta)$ in this situation. In the same principle, after parameter compositions are defined in Definition 4 for $f(\mathbf{x}, \theta) = f(\mathbf{x}, \varphi(\psi(\theta)))$, the lowest parameter size within θ, ψ and φ, will be the upper bound of $f(\mathbf{x}, \theta)$. ■

For Bayesian classifiers defined by eq. (5a), one can rewrite it in a form of:

$$y^* = arg \min Rsik(y, \{\theta_\lambda, \theta_\mathbf{C}\}), \tag{14}$$

where $\theta_\lambda = (\lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{21}, \lambda_{22}, \lambda_{23})$ and $\theta_C = (p(t_1), p(t_2), p(\mathbf{x}|t_1), p(\mathbf{x}|t_2)$ in binary classifications, with $\theta_\lambda \cap \theta_\mathbf{C} = \emptyset$ for their disjoint sets. Let E (or Rej) be the total Bayesian error (or reject) in binary classifications:

$$\begin{array}{l} E(y^*, \theta) = E_1 + E_2 = \\ \qquad \int\limits_{R_2} p(t_1) p(\mathbf{x}|t_1) d\mathbf{x} + \int\limits_{R_1} p(t_2) p(\mathbf{x}|t_2) d\mathbf{x}, \\ Rej(y^*, \theta) = Rej_1 + Rej_2 = \\ \qquad \int\limits_{R_3} p(t_1) p(\mathbf{x}|t_1) d\mathbf{x} + \int\limits_{R_3} p(t_2) p(\mathbf{x}|t_2) d\mathbf{x}. \end{array} \tag{15}$$

Based on eqs. (7) and (12), the total error (or reject) of Bayesian classifiers defined by eq. (15) shows a form of composition of parameters:

$$\begin{array}{r} E(y^*, \{\theta_\lambda, \theta_\mathbf{C}\}) {=} E(y^*, \{\mathbf{x_b}(\mathbf{T_r}(\theta_\lambda)), \theta_\mathbf{C}\}), \\ Rej(y^*, \{\theta_\lambda, \theta_\mathbf{C}\}) {=} Rej(y^*, \{\mathbf{x_b}(\mathbf{T_r}(\theta_\lambda)), \theta_\mathbf{C}\}) \end{array} \tag{16}$$

where $\mathbf{x_b}$ and $\mathbf{T_r}$ are two functions of the parameters. λ_{ij} $(i = 1, 2, j = 1, 2, 3)$ are usually input parameters, but T_{rk} $(k = 1, 2)$ can serve as either intermediate parameters or input ones.

Theorem 2: In abstaining binary classifications, the total number of independent parameters within the cost terms for defining Bayesian classifiers, y^*, should be at most two $(N_{IP}(y^*, \theta) \leq 2)$. Therefore, applications of cost terms of $\theta_\lambda = (\lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{21}, \lambda_{22}, \lambda_{23})$ in the traditional cost sensitive learning will exhibit a parameter redundancy for calculating Bayesian $E(y^*)$ and $Rej(y^*)$ even after assuming $\lambda_{11} = \lambda_{22} = 0$, and $\lambda_{12} = 1$ as the conventional way in classifications [13][27].

Proof: Applying (14) and (13) in Lemma 1, one can have $N_{IP}(y^*, \theta) \leq \min(p = 6, q = 2, r = 4) = 2$ for defining Bayesian classifiers from θ. However, when imposing three constraints on $\lambda_{11} = \lambda_{22} = 0$, and $\lambda_{12} = 1$, θ will provide three free parameters in the cost matrix in a form of:

$$\begin{array}{l} \lambda_{21} {=} \lambda_{21} \\ \lambda_{13} {=} \dfrac{T_{r1}(T_{r2} * \lambda_{21} + T_{r2} - \lambda_{21})}{T_{r1} + T_{r2} - 1} \\ \lambda_{23} {=} \dfrac{T_{r2}(T_{r1} * \lambda_{21} + T_{r1} - \lambda_{21})}{T_{r1} + T_{r2} - 1}, \end{array} \tag{17}$$

which implies a parameter redundancy for calculating Bayesian $E(y^*)$ and $Rej(y^*)$. ■

Remark 1: Theorem 2 describes that Bayesian classifiers with a reject option will suffer a difficulty of uniquely interpreting cost terms. For example, one can even enforce the following two settings:

$$\left\{ \begin{array}{l} \lambda_{11} = 0,\ \lambda_{12} = 1,\ 0 \leq \lambda_{13} \leq 1, \\ \lambda_{21} = 1,\ \lambda_{22} = 0,\ 0 \leq \lambda_{23} \leq 1, \end{array} \right.$$

or

$$\left\{ \begin{array}{l} \lambda_{11} = 0,\ 1 \leq \lambda_{12},\ \lambda_{13} = 1, \\ 1 \leq \lambda_{21},\ \lambda_{22} = 0,\ \lambda_{23} = 1. \end{array} \right.$$

for achieving the same Bayesian classifier, as well as their $E(y^*)$ and $Rej(y^*)$. However, the two sets of settings entail different meanings and do not show the equivalent relations except through eq. (7). Hence, a confusion may be introduced when attempting to understand behaviors of error and reject rates with respects to different sets of cost terms. For this reason, cost terms may present an intrinsic problem for defining a generic form of settings in cost sensitive learning if a reject option is enforced.

Remark 2: While Theorem 2 only shows an estimation of upper-bound of $N_{IP}(y^*, \theta)$ for Bayesian classifiers with a reject option because of missing a closed-form solution of $E(y^*, \theta)$, one can prove on $N_{IP}(y^*, \theta) = 1$ for Bayesian classifiers without rejection. A single independent parameter from the cost terms can be formed as $(\lambda_{12} - \lambda_{11})/(\lambda_{21} - \lambda_{22})$.

Remark 3: We suggest to apply independent parameters for the design and cost analysis of Bayesian classifiers. The total number of independent parameters of $N_{IP}(y^*, \theta)$ is changeable and dependent on the reject option of Bayesian classifiers. If rejection is not considered, we suggest $\theta = (\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = 1, \lambda_{21} > 0)$ for the cost or error sensitivity analysis. A single independent cost parameter, λ_{21}, is capable of governing complete behaviors of error rate. For a reject option, we suggest $\theta = (0 \leq T_{r1}, 0 \leq T_{r2},\ and\ T_{r1} + T_{r1} \leq 1)$ for the cost, error, or reject sensitivity analysis, which will lead to a unique interpretation to the analysis.

C. Examples of Bayesian Classifiers on Univariate Gaussian Distributions

This section will consider abstaining Bayesian classifiers on Gaussian distributions. As a preliminary study, a univariate feature in [4] is adopted for the reason of showing theoretical fundamentals as well as the closed-form solutions. Therefore, if the relevant knowledge of $p(t_i)$ and $p(x|t_i)$ is given, one can depict the plots of $p(t_i|x)$ from calculation of eq. (1) (Fig. 1). Moreover, when λ_{ij} is known, the classification regions of R_1 to R_3 in terms of x_{bj} will be fixed for Bayesian classifiers. After the regions R_1 to R_3, or x_{bj}, are determined, Bayesian risk will be obtained directly. One can see that these boundaries can be obtained from the known data of δ_i when solving an equality equation on (6a) or (6b):

$$\frac{p(x = x_c|t_1)p(t_1)}{P(x = x_c|t_2)p(t_2)} = \delta_i \tag{18}$$

The data of δ_i can be realized either from cost terms λ_{ij}, or from threshold T_{ri} (see eq. (6)). By substituting the exact data of $p(t_i)$ and $p(x|t_1) \sim N(\mu_i, \sigma_i)$ for Gaussian distributions, where μ_i and σ_i represent the mean and standard deviation to the ith class, and the data of δ_i (say, for $\delta_1 = (1 - T_{r1})/T_{r1}$ from the given T_{r1}) into (18), one can obtain the closed-form solutions to the boundary points (say, for x_{b1} and x_{b4}):

$$x_{b1,4} = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2}{\sigma_1^2 - \sigma_2^2} \mp \frac{\sigma_1\sigma_2\sqrt{\alpha}}{\sigma_1^2 - \sigma_2^2},\ if\ \sigma_1 \neq \sigma_2 \tag{19a}$$

$$x_{b1} = \frac{\mu_1 + \mu_2}{2} + \frac{\sigma^2}{\mu_2 - \mu_1} ln(\frac{p(t_1)}{p(t_2)}\frac{1}{\delta_1}),\ if\ \sigma_1 = \sigma_2 = \sigma \tag{19b}$$

where α is an intermediate variable defined by:

$$\alpha = (\mu_1 - \mu_2)^2 - (2\sigma_1^2 - 2\sigma_2^2) ln(\frac{p(t_1)\sigma_2}{p(t_2)\sigma_1}\frac{1}{\delta_1}). \tag{19c}$$

Eq. (19) is also effective for Bayesian classifiers in the case of "*no rejection*". However, only cost terms, $\lambda_{ij}(i, j = 1, 2)$, will define the data of δ_1. The general solution to abstaining classifiers has four boundary points by substituting two threshold T_{r1} and T_{r2}, respectively. For the conditions shown in Fig. 1d, T_{r1} will lead to x_{b1} and x_{b4}, and T_{r2} to x_{b2} and x_{b3}, respectively. Eq. (19a) shows a general form for achieving two boundary points from one data point of δ_1, and eq. (19b) is specific for reaching a single boundary point only when the standard deviations of two classes are the same. Substituting the other data of δ_2 into eq. (19) will yield another pair of data x_{b2} and x_{b3}, or a single one x_{b2}, in a similar form of eq. (19).

Like the solution for boundary points, cross-over point(s) can also be obtained from solving eq. (18) or (19) by substituting $\delta_i = 1$. One can prove that three specific cases will be met with the cross-over point(s) from the solution of eq. (18), namely, two, one, or zero cross-over point(s). The case for the two cross-over points appears only when $\alpha > 0$ in eq. (19c), and two curves of $p(t_1|x)$ and $p(t_2|x)$ demonstrate the non-monotonicity (Fig. 1b) through the equality $p(t_1|x) = 1 - p(t_2|x)$. When the associated standard deviations are equal for the two classes, i.e., $\sigma_1 = \sigma_2$, only one cross-over point appears, which corresponds to the monotonous curves of $p(t_1|x)$ and $p(t_2|x)$ (Fig. 1a). The case for the zero cross-over point occurs when $\alpha < 0$, which corresponds to no real-value (but complex-value) solution to eq. (19a) and to situations of non-monotonous curves of $p(t_1|x)$ and $p(t_2|x)$. In the followings, we will discuss several specific cases for rejections with respect to the cross-over points between the $p(t_1|x)$ and $p(t_2|x)$ curves, as well as to the associated settings on T_r and λ_{ij}. A term is applied to describe every case. For example, "Case_k_BU" indicates "*k*" for the kth case, "B" (or "M") for Bayesian (or mutual-information) classifiers, and "G" (or "U") for Gaussian (or uniform) distributions.

$Case\_1\_BG$: *No rejection.*
For a binary classification, Chow [3] showed that, when $T_{r1} = T_{r2} \geq 0.5$, there exists no rejection for classifiers. The novel constraint of $T_{r1} + T_{r2} \leq 1$ shown in eq. (6e) suggests that the setting should be $T_{r1} = T_{r2} = 0.5$ when the thresholds are the input data. Users need to specify an option for "*no rejection*" or "*rejection*" as an input. When "*no rejection*" is selected, the conventional scheme of cost terms from a two-by-two matrix will be sufficient. Any usage of a two-by-three matrix will introduce some confusion that will be illustrated in the later section by Example 1. In addition, one cannot consider $\lambda_{13} = \lambda_{23} = 0$ as the defaults for the cost matrix in this case.

$Case\_2\_BG$: *Rejection to all or to a complete class.*
In discussing this case, we relax the constraints in eq. (6e) for including the zero values of the thresholds. Chow [3] showed that, whenever $T_r = 0$, a classifier will reject all patterns. Substituting zero values for thresholds into eq. (7), one will obtain solutions for $\lambda_{11} = \lambda_{22} = \lambda_{13} = \lambda_{23} = 0$. These results imply that no cost is received even for a reject decision to a pattern. Obviously, a case like this should be avoided. In some situations, if one intends to reject a complete class (say, Class 1), its associated cost terms should be set to zero (say, $\lambda_{11} = \lambda_{13} = 0$). We call these situations as "*one-class and reject-class*" classification, since only two categories are identified, that is, "Class 2" and "Reject Class", respectively.

$Case\_3\_BG$: *Rejection in two cross-over points x_{c1} and x_{c2}.*
The necessary condition for realizing this case is derived from

TABLE I
REJECTION SETTINGS FOR BAYESIAN CLASSIFIERS IN UNIVARIATE GAUSSIAN DISTRIBUTIONS
$(x_{b1} < x_c < x_{b2} \quad or \quad x_{b1} < x_{c1} < x_{b2} < x_{b3} < x_{c2} < x_{b4})$

Cross-over Point(s) (Reference Figure)	Rejection Thresholds	Reject region(s)	Remarks
Two (Fig. 1d)	$T_{r1} = 0.5$, $T_{r2} = 0.5$	$\emptyset$	No Rejection
	$T_{r1} \geq 0.5$, $1 - \max(p(t_2\|x)) < T_{r2} < 0.5$	$[x_{c1}, x_{b2})$ and $(x_{b3}, x_{c2}]$	-
	$T_{r1} < 0.5$, $T_{r2} \geq 0.5$	$[x_{b1}, x_{c1})$ and $(x_{c2}, x_{b4}]$	-
	$T_{r1} < 0.5$, $1 - \max(p(t_2\|x)) < T_{r2} < 0.5$	$[x_{b1}, x_{b2})$ and $(x_{b3}, x_{b4}]$	General Rejection
	$T_{r1} < 0.5$, $T_{r2} < 1 - \max(p(t_2\|x))$	$[x_{b1}, x_{b4}]$	"*Class-1 and Reject-class*" Classification
	$T_{r1} = 0$, $T_{r2} < 1$	$(-\infty, x_{b2})$ and (x_{b3}, ∞)	"*Class-2 and Reject-class*" Classification
One (Fig. 1c)	$T_{r1} = 0.5$, $T_{r2} = 0.5$	$\emptyset$	No Rejection
	$T_{r1} \geq 0.5$, $T_{r2} < 0.5$	$[x_c, x_{b2})$	-
	$T_{r1} < 0.5$, $T_{r2} \geq 0.5$	$[x_{b1}, x_c)$	-
	$T_{r1} < 0.5$, $T_{r2} < 0.5$	$[x_{b1}, x_{b2})$	General Rejection
Zero (Fig. 1d)	$T_{r1} \geq 1 - \min(p(t_1\|x))$	$\emptyset$	"*Majority-taking-all*" Classification
	$T_{r1} < 1 - \min(p(t_1\|x))$ $T_{r2} < 1 - \max(p(t_2\|x))$	$[x_{b1}, x_{b4}]$	"*Majority-class and Reject-class*" Classification
	$T_{r1} < 1 - \min(p(t_1\|x))$ $T_{r2} > 1 - \max(p(t_2\|x))$	$[x_{b1}, x_{b2})$ and $(x_{b3}, x_{b4}]$	General Rejection
	$T_{r1} = 0$ $T_{r2} > 1 - \max(p(t_2\|x)) > 0.5$	$(-\infty, x_{b2})$ and (x_{b3}, ∞)	"*Minority-class and Reject-class*" Classification
Zero, one and Two (Fig.1)	$T_{r1} = T_{r2} = 0$	$(-\infty, \infty)$	Rejection to All

eq. (18) for $\alpha > 0$ while assuming $\delta_i = 1$:

$$\frac{\lambda_{12} - \lambda_{11}}{\lambda_{21} - \lambda_{22}} < \frac{p(t_2)\sigma_1}{p(t_1)\sigma_2} e^{\frac{\mu_1 - \mu_2}{2(\sigma_1^2 - \sigma_2^2)}} \tag{20}$$

The general situation within this case is when $T_{r1} < 0.5$ and $1 - \max(p(t_2|x)) < T_{r2} < 0.5$, in which the reject region R_3 is divided by two ranges. When $T_{r1} < 0.5$ and $T_{r2} < 1 - \max(p(t_2|x)) < 0.5$, only one class is identified, but all other patterns are classified into a reject class. Therefore, we refer this situation as "*Class 1 and Reject-class*" classification. Table I also lists the other situations for the rejections from the different settings on T_{rj}.

$Case\_4\_BG$: *Rejection in one cross-over point* x_c.
The general condition for realizing this case in the context of classifications is not based from setting an equality condition on (20) for $\alpha = 0$. We neglect such setting in this case, but assign it into Case_5_BG. As demonstrated in eq. (19b), the general condition of this case is a simply setting $\sigma_1 = \sigma_2$. Since the monotonicity property is enabled for the curves of $p(t_1|x)$ and $p(t_2|x)$ in this case, a single reject region is formed (Fig. 1c).

$Case\_5\_BG$: *Rejection in zero cross-over point.*
The general condition for realizing this case corresponds to a violation of the criterion on (19a), or $\alpha < 0$ in (20). In this case, one class always shows a higher value of the posterior probability distribution over the other one in the whole domain of x. From definitions in the study of class imbalanced dataset [14] [16], if $p(t_1) > p(t_2)$ in binary classifications, Class 1 will be called a "*majority*" class and Class 2 a "*minority*" class. Supposing that $p(t_1|x) > p(t_2|x)$, when $T_{r1} > 1 - \min(p(t_1|x))$, all patterns will be considered as Class 1. We call these situations as a "*Majority-taking-all*" classification. Due to the constraints like $T_{r1} + T_{r2} \leq 1$ and $p(t_1|x) + p(t_2|x) = 1$, one is unable to realize a "*Minority-taking-all*" classification. When $T_{r1} < 1 - \min(p(t_1|x))$ and $T_{r2} < 1 - \max(p(t_2|x))$, all patterns will be partitioned into one of two classes, that is, majority and rejection. We call these situations "*Majority-class and Reject-class*" classifications. The situations of "*Minority-class and Reject-class*" classification occur if $T_{r2} > 1 - \max(p(t_2|x)) > 0.5$ and $T_{r1} = 0$.

Since the study of imbalanced data learning received more attentions recently [16][17][18], one related theorem of Bayesian classifiers is derived below for elucidating their important features.

Theorem 3: Consider a binary classification with an exact knowledge of one-dimensional Gaussian distributions. If a zero-one cost function is applied, Bayesian classifiers without rejection will satisfy the following rule:

$$\begin{aligned} if\ p_{min} = \min(p(t_1), p(t_2)) \to 0,\ and \\ \lambda_{11} = \lambda_{22} = 0, \lambda_{12} = \lambda_{21} = 1 \\ then\ E \to E_{max} = p_{min}, \end{aligned} \tag{21}$$

which indicates that the classifiers have a tendency of reaching the maximum Bayesian error, E_{max}, by misclassifying all rare-class patterns in imbalanced data learning.

Proof: We will prove the misclassification of all rare-class patterns first. Suppose $p(t_2)$ represents the prior probability of the "*minority*" or "*rare*" class in imbalanced data learning and consider the special case firstly on the equal variances for two classes (Fig. 1a). When $p(t_2)$ approaches to zero, x_c will approach infinity from using eq. (19b) with $\delta_i = 1$. This result indicates that Bayesian classifiers will assign all patterns into the "*majority*" class in classifications. When the variances are

not equal, eqs. (19a) and (19c) with $\delta_i = 1$ will be applicable (Fig. 1b). One can obtain the relation $\alpha < 0$ for the case that no cross-over point occurs on $p(t_i|x)$ plots when $p(t_2)$ approaches to zero. Only the "*majority*" class is identified from using Bayesian classifiers in this case. The equality of $E_{max} = p_{min}$ suggests an upper bound of Bayesian error (See Appendix B). If violating this bound, Bayesian classifiers will adjust themselves for achieving the smallest error rate. ■

D. Examples of Bayesian Classifiers on Univariate Uniform Distributions

Chow [3] presented a study on rejection from Bayesian classifiers along uniform distributions for one-dimensional problems. This section will extend Chow's results by providing general formulas of parameterized distributions. A binary classification is considered. The two uniform distributions on two classes are given:

$$p(x|t_1) = \begin{cases} \dfrac{1}{x_2 - x_1} & when\ x_1 \le x \le x_2 \\ 0 & otherwise \end{cases} \tag{22a}$$

$$p(x|t_2) = \begin{cases} \dfrac{1}{x_4 - x_3} & when\ x_3 \le x \le x_4 \\ 0 & otherwise \end{cases} \tag{22b}$$

Three specific cases, shown in Fig. 2, will appear, namely, "*Partially overlapping*", "*Fully overlapping by one class*", and "*Separating*" between two distributions for eq. (22). We will discuss each case with respect to their rejection settings.

$Case\_1\_BU$: *Partially overlapping between two distributions.*

Suppose that the constraints for this case are:

$$x_1 \le x \le x_4,\ and\ x_1 \le x_3 \le x_2 \le x_4. \tag{23}$$

When the relevant knowledge of $p(t_i)$ and $p(x|t_i)$ is given, one is able to gain the posterior probabilities from eqs. (1) and (21) by a closed form:

$$p(t_1|x) = \begin{cases} 1 & when\ x_1 \le x < x_3 \\ \dfrac{p(t_1)(x_4 - x_3)}{p(t_1)(x_4 - x_3) + p(t_2)(x_2 - x_1)} & when\ x_3 \le x \le x_2 \\ 0 & otherwise \end{cases} \tag{24a}$$

$$p(t_2|x) = \begin{cases} 1 & when\ x_2 \le x < x_4 \\ \dfrac{p(t_2)(x_2 - x_1)}{p(t_1)(x_4 - x_3) + p(t_2)(x_2 - x_1)} & when\ x_3 \le x \le x_2 \\ 0 & otherwise \end{cases} \tag{24b}$$

Based on the Bayesian rules of eq. (10) and eq. (24), one can immediately determine $R_1 = [x_1, x_3)$ and $R_2 = [x_2, x_4]$ directly for Class 1 and Class 2, respectively, as shown in Fig. 2. The remaining range is denoted as $R_i = [x_3, x_2)$, since it needs to be identified further depending on the thresholds defined in (7). Due to the simplicity of the uniform distributions, one is able to realize analytical solutions directly for Bayesian classifiers. The probabilities of errors and rejects are calculated from :

$$E = \begin{cases} \dfrac{p(t_2)(x_2 - x_3)}{(x_4 - x_3)}, & if\ f(x \in R_i) = y_1 \\ \dfrac{p(t_1)(x_2 - x_3)}{(x_2 - x_1)}, & if\ f(x \in R_i) = y_2 \\ 0, & if\ f(x \in R_i) = y_3 \end{cases} \tag{25}$$

and

$$Rej = \begin{cases} 0, & if\ f(x \in R_i) = y_1 \\ 0, & if\ f(x \in R_i) = y_2 \\ (x_2 - x_3)[\dfrac{p(t_1)}{(x_2 - x_1)} + \dfrac{p(t_2)}{(x_4 - x_3)}], & \\ \quad if\ f(x \in R_i) = y_3 \end{cases} \tag{26}$$

We use $f(x \in R_i) = y_j$ to describe a decision that R_i is a range of Class j. Eq. (25) demonstrates that Bayesian classifiers with uniform distributions of classes will receive error either from Class 1 or from Class 2, but not both. When setting cost terms properly, zero error can be achieved with conditions of rejection on both classes as shown in eq. (26). It is interesting to observe that cost terms can only control the error types or give the appearance of rejection, but not the degree of them. This is significantly different from Bayesian classifiers with Gaussian distributions of classes.

$Case\_2\_BU$: *Fully overlapping by one class.*

The constraints for this case are:

$$x_1 \le x \le x_2,\ and\ x_1 \le x_3 \le x_4 \le x_2. \tag{27}$$

and the posterior probabilities are:

$$p(t_1|x) = \begin{cases} 1 & when\ x_1 \le x < x_3 \\ & or\ x_4 \le x < x_2 \\ \dfrac{p(t_1)(x_4 - x_3)}{p(t_1)(x_4 - x_3) + p(t_2)(x_2 - x_1)} & when\ x_3 \le x \le x_4 \\ 0 & otherwise \end{cases} \tag{28a}$$

$$p(t_2|x) = \begin{cases} \dfrac{p(t_2)(x_2 - x_1)}{p(t_1)(x_4 - x_3) + p(t_2)(x_2 - x_1)} & when\ x_3 \le x \le x_4 \\ 0 & otherwise \end{cases} \tag{28b}$$

Following the similar way in the previous case, one can obtain the analytical results:

$$E = \begin{cases} p(t_2), & if\ f(x \in R_i) = y_1 \\ \dfrac{p(t_1)(x_4 - x_3)}{(x_2 - x_1)}, & if\ f(x \in R_i) = y_2\ , \\ 0, & if\ f(x \in R_i) = y_3 \end{cases} \tag{29}$$

and

$$Rej = \begin{cases} 0, & if\ f(x \in R_i) = y_1 \\ 0, & if\ f(x \in R_i) = y_2 \\ p(t_1)\dfrac{(x_4 - x_3)}{x_2 - x_1} + p(t_2), & if\ f(x \in R_i) = y_3 \end{cases} . \tag{30}$$

Specific solutions will be received in this case on Class 2, which is full overlapped within Class 1. All patterns within

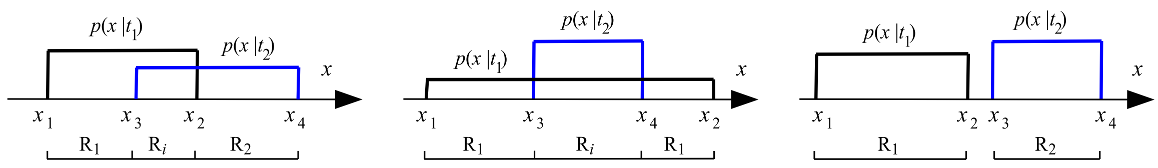


Fig. 2. Classification scenarios for univariate uniform distributions.

Class 2 may be misclassified or rejected fully depending on the settings of cost terms.

$Case\_3\_BU$: *Separation between two distributions.*
One is able to obtain the exact solutions without any error and reject. Cost terms are useless in this case.

III. Mutual-information based Classifiers with a Reject Option

A. Mutual-information based Classifiers

Definition 5 (Mutual-information classifier): A mutual-information classifier is the classifier which is obtained from the maximization of mutual information over all patterns:

$$y^+ = arg \max_y NI(T = t, Y = y), \tag{31}$$

where T and Y are the target variable and decision output variable, t and y are their values, respectively. For simplicity, we denote $NI(T = t, Y = y) = NI(T, Y)$ as the normalized mutual information in a form of [11]:

$$NI(T,Y) = \frac{I(T,Y)}{H(T)} \tag{32a}$$

where $H(T)$ is the entropy based on the Shannon definition [37] to the target variable,

$$H(T) = -\sum_{i=1}^{m} p(t_i) log_2 p(t_i), \tag{32b}$$

and $I(T,Y)$ is mutual information between two variables of T and Y [38]:

$$I(T,Y) = \sum_{i=1}^{m}\sum_{j=1}^{m+1} p(t_i,y_j) log_2 \frac{p(t_i,y_j)}{p(t_i)p(y_j)}, \tag{32c}$$

where m is a total number of classes in T. For binary classifications, we set $m = 2$. In (32), $p(t,y)$ is the joint distribution between the two variables, and $p(t)$ and $p(y)$ are the marginal distributions which can be derived from [38]:

$$p(t) = \sum_y p(t,y), \ \ and \ \ p(y) = \sum_t p(t,y). \tag{33}$$

Mathematically, eq. (31) expresses that y^+ is an optimal classifier in terms of the maximal mutual information, or relative entropy, between the target variable T and decision output variable Y. The physical interpretation of relative entropy is a measurer of probability similarity between the two variables. Note that the present definition of NI is asymmetry to the variables T and Y for the normalization term of $H(T)$ (=constant, for given $p(t)$), but will not make a difference for arriving at the optimal y^+ defined by (31). We adopt Shannon's definition of entropy for the reason that no free parameter is introduced. A normalization scheme is applied so that a relative comparison can be made easily among classifiers.

Definition 6 (Augmented Confusion Matrix [11]): An augmented confusion matrix will include one column for a rejected class, which is added on a conventional confusion matrix:

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} & c_{1(m+1)} \\ c_{21} & c_{22} & \cdots & c_{2m} & c_{2(m+1)} \\ & & \cdots & & \\ c_{m1} & c_{m2} & \cdots & c_{mm} & c_{m(m+1)} \end{bmatrix}, \tag{34}$$

where c_{ij} represents the number of the ith class that is classified as the jth class. The row data corresponds to the exact classes, and the column data corresponds to the prediction classes. The last column represents a reject class. The relations and constraints of an augmented confusion matrix are:

$$C_i = \sum_{j=1}^{m+1} c_{ij}, \ C_i > 0, \ c_{ij} \geq 0, \ i = 1, 2, \cdots, m \tag{35}$$

where C_i is the total number for the ith class. The data for C_i is known in classification problems.

In this work, supposing that the input data for classifications are exactly known about the prior probability $p(t_i)$ and the conditional probability density function $p(\mathbf{x}|t_i)$, one is able to derive the joint distribution matrix in association with the confusion matrix:

$$\begin{aligned} & p_{ij} = p(t_i, y_j) = \int_{R_j} p(t_i)p(\mathbf{x}|t_i)d\mathbf{x} \approx \frac{c_{ij}}{n} = p_e(t_i, y_j), \\ & i = 1, 2, \cdots, m, \ j = 1, 2, \cdots, m+1 \end{aligned} \tag{36}$$

where R_j is denoted as the region in which every pattern $\mathbf{x}$ is identified as the jth class, and $p_e(t_i, y_j)$ is the empirical probability density for applications where only a confusion matrix is given. In those applications, the total number of patterns n is generally known.

Eq. (36) describes the approximation relations between the joint distribution and confusion matrix. If the knowledge about $p(t_i)$ and $p(\mathbf{x}|t_i)$ are exactly known, one can design a mutual information classifier directly. If no initial information is known about $p(t_i)$ and $p(\mathbf{x}|t_i)$, the empirical probability density of joint distribution, $p_e(t_i, y_j)$, can be estimated from the confusion matrix [11]. This treatment, based on the frequency principle of a confusion matrix, is not mathematically rigorous, but will offer a simple approach for classifiers to apply the entropy principle for wider applications.

$$p(t,y)=\begin{bmatrix} \frac{p(t_1)}{2}[1-erf(X_{11})] & \frac{p(t_1)}{2}[1-erf(X_{12})] & \frac{p(t_1)}{2}[erf(X_{11})+erf(X_{12})] \\ \frac{p(t_2)}{2}[1-erf(X_{21})] & \frac{p(t_2)}{2}[1-erf(X_{22})] & \frac{p(t_2)}{2}[erf(X_{21})+erf(X_{22})] \end{bmatrix}, \tag{42a}$$

Considering binary classifications, one will have the following formula for the joint distribution $p(t, y)$:

$$p(t,y)= \begin{bmatrix} \int\limits_{R_1} p(t_1)p(\mathbf{x}|t_1)d\mathbf{x} & \int\limits_{R_2} p(t_1)p(\mathbf{x}|t_1)d\mathbf{x} & \int\limits_{R_3} p(t_1)p(\mathbf{x}|t_1)d\mathbf{x} \\ \int\limits_{R_1} p(t_2)p(\mathbf{x}|t_2)d\mathbf{x} & \int\limits_{R_2} p(t_2)p(\mathbf{x}|t_2)d\mathbf{x} & \int\limits_{R_3} p(t_2)p(\mathbf{x}|t_2)d\mathbf{x} \end{bmatrix}. \tag{37}$$

The marginal distribution for $p(t)$ is in fact the given information of prior knowledge about the classes:

$$p(t) = (p(t_1), p(t_2))^T \tag{38}$$

where the superscript "T" represents a transpose, and the marginal distribution for $p(y)$ is:

$$\begin{aligned} &p(y) = (p(y_1), p(y_2), p(y_3)) = (\int\limits_{R_1} Qd\mathbf{x}, \int\limits_{R_2} Qd\mathbf{x}, \int\limits_{R_3} Qd\mathbf{x}) \\ &Q = p(t_1)p(\mathbf{x}|t_1) + p(t_2)p(\mathbf{x}|t_2). \end{aligned} \tag{39}$$

Substituting (37) - (38) into (32), one can obtain the formula of NI in terms of $p(t_i)$ and $p(\mathbf{x}|t_i)$. When the prior knowledge of $p(t_i)$ is given, the conditional entropy $H(T)$ in eq. (32b) will be unchanged during classifier learnings. This is why we use this term to normalize the mutual information in (32a).

B. Examples of Mutual-information Classifiers on Univariate Gaussian Distributions

Mutual-information classifiers, like Bayesian classifiers, also provide a general formulation to classifications. They are able to process classifications with or without rejection. This section will aim at deriving novel formulas necessary for the design and analysis of mutual-information classifiers under assumptions of Gaussian distributions. The assumptions, or given input data, for the derivations are kept the same as those for Bayesian classifiers shown in Section II, except that cost terms of λ_{ij} are not given as the input, but will be displayed as the output of the classifiers. In other words, mutual information classifiers will automatically calculate the two thresholds that can lead to the cost terms through eq. (7). However, due to a redundancy among six cost terms, one will fail to obtain the unique solution of the cost terms, which is demonstrated in Example 1 of Section IV.

Generally, one is unable to derive a closed-form solution to mutual-information classifiers. One of the obstacles is the nonlinear complexity of solving error functions. Therefore, this work only provides semi-analytical solutions for mutual information classifiers. When substituting $p(t_i)$ and $p(x|t_i)$ into eqs. (31) and (32), one will encounter the process of solving an inverse problem on the following function:

$$\max_{y \in Y} NI(T,Y) = \max f(x, \theta = (p(t_i), p(x|t_i), x_{bj})), \tag{40}$$

for searching the boundary points x_{bj} from error functions. Only numerical solutions can be obtained for x_{bj}, except for a special case. Whenever a reject option is set, mutual-information classifiers will generate classification regions, R_i $(i = 1, 2, 3)$, automatically according to the given data of $p(t_i)$ and $p(x|t_i)$, as shown in Table II. In the followings, some specific cases of mutual-information classifiers will be discussed in related to a reject option.

$Case\_1\_MG$: *No rejection in one cross-over point* x_c *when* $p(t_1) = p(t_2)$ *and* $\sigma_1 = \sigma_2$.

This is a very special case where one is able to obtain a closed-form solution to mutual-information classifiers. Under the conditions of $p(t_1) = p(t_2)$, $\sigma_1 = \sigma_2$, and two by two joint distribution matrix for no rejection, one can get a single boundary point x_b, coincident to the cross-over point x_c, for partitioning the classification regions:

$$\begin{aligned} &x_b = x_c = \frac{\mu_1+\mu_2}{2}, \\ &if\ \mu_1 < \mu_2\ then\ R_1 = (-\infty, x_b), R_2 = [x_b, \infty), R_3 = \emptyset. \end{aligned} \tag{41}$$

This result exhibits similar results for Bayesian classifiers, which leads to the same error values between the two types of classifiers. Therefore, eq. (41) indicates that $y^+ = y^*$ to be fully equivalent between mutual-information classifiers and Bayesian classifiers under the conditions of $p(t_1) = p(t_2)$ and $\sigma_1 = \sigma_2$ when no reject option is selected.

$Case\_2\_MG$: *Rejection in one cross-over point* x_c *and* $\sigma_1 = \sigma_2$.

When we relax the condition in the case above on $p(t_1) \neq (t_2)$ and with a reject option, the solutions to mutual-information classifiers become not fully analytical. The key step for missing such an analytical solution comes from a determination of x_{bj}. In this case, due to the condition that $\sigma_1 = \sigma_2$, one will have a single cross-over point x_c as the general case in binary classifications for Gaussian distributions. If a reject option is selected, one will generally have two boundary points x_{b1} and x_{b2}. Suppose $\mu_1 < \mu_2$ and $x_{b1} < x_{b2}$, one can partition classification regions as: $R_1 = (-\infty, x_{b1}), R_2 = [x_{b2}, \infty)$, and $R_3 = [x_{b1}, x_{b2})$. Supposing the two boundary points are given, one can have a closed-form formula on eq. (37):

(Please see the equation on the top of this page)

where $erf(\cdot)$ is an error function, and

$$X_{ij} = \frac{\mu_i - x_{bj}}{\sqrt{2}\sigma_i},\ i=1,2,\ j=1,2. \tag{42b}$$

In this work, we adopt a numerical approach to search the results on x_{b1} and x_{b2}. Whenever these values are known, one can get the error rate and reject rate from:

$$\begin{aligned} E &= E_1 + E_2 = p(t_i=1, y_j=2) + p(t_i=2, y_j=1) \\ &= \frac{p(t_1)}{2}[1-erf(X_{12})] + \frac{p(t_2)}{2}[1-erf(X_{21})] \end{aligned} \tag{43a}$$

TABLE II
CLASSIFICATION REGIONS FOR MUTUAL-INFORMATION CLASSIFIERS IN UNIVARIATE GAUSSIAN DISTRIBUTIONS OF FIG. 1 ($x_{b1} < x_{b2} < x_{b3} < x_{b4}$)

Reject Option	Cross-over Point(s)	Boundary Point(s)	Class of R_1	Class of R_2	Class of R_3
No Rejection	x_c	x_b	$(-\infty, x_b)$	$[x_b, \infty)$	$\emptyset$
	x_{c1}, x_{c2}	x_{b1}, x_{b2}	$(-\infty, x_{b1})$ and (x_{b2}, ∞)	$[x_{b1}, x_{b2}]$	$\emptyset$
Rejection	x_c	x_{b1}, x_{b2}	$(-\infty, x_{b1})$	$[x_{b2}, \infty)$	$[x_{b1}, x_{b2})$
	x_{c1}, x_{c2}	$x_{b1}, x_{b2}, x_{b3}, x_{b4}$	$(-\infty, x_{b1})$ and (x_{b2}, ∞)	$[x_{b2}, x_{b3}]$	$[x_{b1}, x_{b2})$ and $(x_{b3}, x_{b4}]$

$$\begin{aligned} Rej &= Rej_1 + Rej_2 \\ &= p(t_i = 1, y_j = 3) + p(t_i = 2, y_j = 3) \\ &= \frac{p(t_1)}{2}[erf(X_{11}) + erf(X_{12})] \\ &\quad + \frac{p(t_2)}{2}[erf(X_{21}) + erf(X_{22})] \end{aligned} \tag{43b}$$

$Case\_3\_MG$: *Rejection in two cross-over points.*
This is a general case for mutual-information classifiers in which four boundary points, x_{bj}, are formed. When the four points obtained numerically during solving eq. (31), the classification regions R_1 to R_3 will be set as shown in Table II. With the condition of $x_{b1} < x_{b2} < x_{b3} < x_{b4}$, the closed-form solution of $p(t, y)$ can be given in a similar way of eq. (42). Additionally, both error and reject rates can be evaluated from $p(t, y)$. For comparing with Bayesian classifiers, the equivalent rejection thresholds are derived from the given data of x_{bj}:

$$\begin{aligned} T_{r1} &= 1 - p(t_1|x = x_{b1}) \\ &= 1 - \frac{p(t_1)\sigma_2 e^{\frac{-(x_{b1}-\mu_1)^2}{2\sigma_1^2}}}{p(t_1)\sigma_2 e^{\frac{-(x_{b1}-\mu_1)^2}{2\sigma_1^2}} + p(t_2)\sigma_1 e^{\frac{-(x_{b1}-\mu_2)^2}{2\sigma_2^2}}} \end{aligned} \tag{44a}$$

$$\begin{aligned} T_{r2} &= 1 - p(t_2|x = x_{b2}) \\ &= 1 - \frac{p(t_2)\sigma_1 e^{\frac{-(x_{b2}-\mu_2)^2}{2\sigma_2^2}}}{p(t_1)\sigma_2 e^{\frac{-(x_{b2}-\mu_1)^2}{2\sigma_1^2}} + p(t_2)\sigma_1 e^{\frac{-(x_{b2}-\mu_2)^2}{2\sigma_2^2}}} \end{aligned} \tag{44b}$$

With the condition of $x_{b1} < x_{b2} < x_{b3} < x_{b4}$ shown in Fig. 1d, substituting either x_{b1} or x_{b4} into (44) will give the same value on T_{r1}, and a similar one for x_{b2} or x_{b3} on T_{r2}. The results of T_{r1} and T_{r2} indicate that mutual-information classifiers will automatically search the rejection thresholds for balancing the error rate and reject rate for the given data of classes. This specific feature will be discussed in Section IV.

C. Examples of Mutual-information Classifiers on Univariate Uniform Distributions

When comparing with Bayesian classifiers, we examine mutual-information classifiers on uniform distributions in this section. The two classes and their conditional probability density functions are given in (22). Three cases will be discussed below.

$Case\_1\_MU$: *Partially overlapping between two distributions.*
In this case (Fig. 2a), one needs to construct joint distribution $p(t, y)$ first. For binary classifiers, $p(t, y)$ is given in the following forms:

$$p(t,y) = \begin{bmatrix} p(t_1) & 0 & 0 \\ \frac{p(t_2)(x_2 - x_3)}{(x_4 - x_3)} & \frac{p(t_2)(x_4 - x_2)}{(x_4 - x_3)} & 0 \end{bmatrix}, \quad if\ f(x \in R_i) = y_1 \tag{45a}$$

$$p(t,y) = \begin{bmatrix} \frac{p(t_1)(x_3 - x_1)}{(x_2 - x_1)} & \frac{p(t_1)(x_2 - x_3)}{(x_2 - x_1)} & 0 \\ 0 & p(t_2) & 0 \end{bmatrix}, \quad if\ f(x \in R_i) = y_2 \tag{45b}$$

$$p(t,y) = \begin{bmatrix} \frac{p(t_1)(x_3 - x_1)}{(x_2 - x_1)} & 0 & \frac{p(t_1)(x_2 - x_3)}{(x_2 - x_1)} \\ 0 & \frac{p(t_2)(x_4 - x_2)}{(x_4 - x_3)} & \frac{p(t_2)(x_2 - x_3)}{(x_4 - x_3)} \end{bmatrix}, \quad if\ f(x \in R_i) = y_3 \tag{45c}$$

Eq. (45) demonstrates three sets of $p(t, y)$ due to diffident decisions may be involved with R_i in Fig. 2a. Substituting (45) into (32), one will obtain three sets of NI's. The closed-form solutions about the decision can be given, but this work adopts a numerical approach for omitting tedious descriptions of the formulas.

$Case\_2\_MU$: *Fully overlapping by one class.*
The formula for $p(t, y)$ in this case (Fig. 2b) is:

$$p(t,y) = \begin{bmatrix} p(t_1) & 0 & 0 \\ p(t_2) & 0 & 0 \end{bmatrix}, \ if\ f(x \in R_i) = y_1 \tag{46a}$$

$$p(t,y) = \begin{bmatrix} \frac{p(t_1)(x_2 - x_1 - x_4 + x_3)}{(x_2 - x_1)} & \frac{p(t_1)(x_4 - x_3)}{(x_2 - x_1)} & 0 \\ 0 & p(t_2) & 0 \end{bmatrix}, \quad if\ f(x \in R_i) = y_2 \tag{46b}$$

$$p(t,y) = \begin{bmatrix} \frac{p(t_1)(x_2 - x_1 - x_4 + x_3)}{(x_2 - x_1)} & 0 & \frac{p(t_1)(x_4 - x_3)}{(x_2 - x_1)} \\ 0 & 0 & p(t_2) \end{bmatrix}, \quad if\ f(x \in R_i) = y_3 \tag{46c}$$

One can get the following results through substituting (46) into (32):

$$NI(t, y) = 0, \ if\ f(x \in R_i) = y_1 \tag{47a}$$

$$0 < NI(t, y) < 1, \ if \ f(x \in R_i) = y_2 (or \ y_3). \quad (47b)$$

Eq. (47a) suggests that the decision for $f(x \in R_i) = y_1$ will produce zero information. Therefore, mutual information classifiers will never make this kind of decisions (but Bayesian classifiers may do so).

$Case\_3\_MU$: *Separation between two distributions.* Mutual-information classifiers will show the perfect solutions as those for Bayesian classifiers.

IV. Comparisons between Bayesian Classifiers and Mutual-information Classifiers

A. General Comparisons

Mutual-information classifiers provide users a wider perspective in processing classification problems, hence a larger toolbox in their applications. For discovering new features in this approach, the present section will discuss general aspects of mutual-information and Bayesian classifiers at the same time for a systematic comparison. The main objective of the comparative study is to reveal their corresponding advantages and disadvantages. Meanwhile, their associated issues, or new challenges, are also presented from the personal viewpoint of the author.

First, both types of classifiers share the same assumptions of requiring the exact knowledge about class distributions and specifying the status of the reject option (Table III). The "*exact knowledge*" feature imposes the most weakness on the two approaches in applications. In other words, the approaches are more theoretically meaningful, rather than directly useful for solving real-world problems. When the exact knowledge is not available, the existing estimation approaches to class distributions [4][33][40] for Bayesian classifiers will be feasible for implementing mutual-information classifiers. The learning targets of Bayesian classifiers involve evaluations of risks or errors, which is mostly compatible with classification goals in real-life applications. However, the concept of mutual information, or entropy-based criteria, is not a common concern or requirement from most classifier designers and users [11].

Second, Bayesian classifiers will ask (or implicitly apply) cost terms for their designs. This requirement provides both advantages and disadvantages depending on applications. The main advantage is its flexibility in offering objective or subjective designs of classifiers. When the exact knowledge is available and reliable, inputing such data will be very simple and meaningful for realizing objective designs. At the same time, subjective designs will always be possible. The main disadvantage may occur for objective designs if one has incomplete information about cost terms. Generally, cost terms are more liable to subjectivity than prior probabilities. In this case, avoiding the introduction of subjectivity is not an easy task for Bayesian classifiers. Mutual-information classifiers, without requiring cost terms, will fall into an objective approach. They carry an intrinsic feature of "*letting the data speak for itself*", which exhibits a significant difference from a subjective version of Bayesian classifiers. However, the current definition of mutual-information classifiers needs to be extended for carrying the flexibility of subjective designs, which is technically feasible by introducing free parameters, such as fuzzy entropy [42].

Third, one of the problems for the current learning targets of Bayesian classifiers is their failure to obtain the optimal rejection threshold in classifications. Although Chow [3] and Ha [22] suggested formulas respectively in forms of:

$$\min Risk(T_r) = E(T_r) + T_r Rej(T_r), \quad (48)$$

or

$$\min \frac{E(T_r)}{Rej(T_r)}, \quad (49)$$

respectively, a minimization from both formulas will lead to a solution of $T_r = 0$ for $Risk = 0$, which implies a rejection of all patterns. Therefore, we can expect to establish a meaningful learning target which is applicable to Bayesian classifiers for determining optimal rejection thresholds. On the contrary, mutual-information classifiers are able to achieve the optimal rejection thresholds as the classifiers' outcomes. The remaining issue is to study the optimal cases in a systematic way.

Fourth, Bayesian classifiers generally fail to handle class imbalanced data properly if no cost terms are specified in classifications, as described in Theorem 3. When one class approximates a smaller (or zero) population and no distinction is made among error types, Bayesian classifiers have a tendency to put all patterns of the smaller class into error, and its NI will be approximately zero, which represents that no information is obtained from classifiers [9]. Mutual-information classifiers display particular advantages in these situations, including cases for abstaining classifications. They provide a solution of balancing error types and reject types without using cost terms. The challenge lies in their theoretical derivation of response behaviors, such as, upper bound and lower bound of $E_i/p(t_i)$ for mutual-information classifiers.

Fifth, mutual-information classifiers will add extra computational complexities and costs over Bayesian classifiers. Both types of classifiers require computations of posterior probability. When these data are obtained, Bayesian classifiers will produce decision results directly. However, mutual-information classifiers will need further procedures, such as, to form a confusion matrix (or a joint distribution matrix), to evaluate NI in (31), and to search boundary points from a non-convex space NI in (40). These procedures will introduce significantly analytical and computational difficulties to mutual-information classifiers, particularly in multiple-class problems with high dimensions.

Note that the discussions above provide a preliminary answer to the question posed in the title of this paper. In another connection, Appendix B presents the tighter bounds between conditional entropy and Bayesian error in binary classifications. Further investigations are expected to search other differences under various assumptions or backgrounds, such as distributions of mixture models, multiple-class classifications in high dimension variables, rejection to a subset of classes [22], and experimental studies from real-world datasets.

TABLE III
DATA INFORMATION FOR BAYESIAN AND MUTUAL-INFORMATION CLASSIFIERS IN BINARY CLASSIFICATIONS

Classifier Type	Required Input		Learning Target	Output Data
	On Data	On Rejection		
Bayesian	$p(t_1)$, $p(t_2)$ $p(\mathbf{x}\|t_1)$, $p(\mathbf{x}\|t_2)$ λ_{11}, λ_{12}, λ_{13} λ_{21}, λ_{22}, λ_{23} (or T_{r1}, and T_{r2})	No or Yes	$\min Risk(y)$ or $\min E(y)$	E_1, E_2, Rej_1, Rej_2, $Risk$, R_1, R_2, R_3, T_{r1}, T_{r2}, ($\{\lambda_{21}/\lambda_{12}\}$, or $\{\lambda_{21}, \lambda_{13}, \lambda_{23}\}$)
Mutual-Information	$p(t_1)$, $p(t_2)$ $p(\mathbf{x}\|t_1)$, $p(\mathbf{x}\|t_2)$	No or Yes	$\max NI(T, Y)$	E_1, E_2, Rej_1, Rej_2, NI, R_1, R_2, R_3, T_{r1}, T_{r2}, ($\{\lambda_{21}/\lambda_{12}\}$, or $\{\lambda_{21}, \lambda_{13}, \lambda_{23}\}$)

B. *Comparisons on Univariate Gaussian Distributions*

Gaussian distributions are important not only in theoretical sense. To a large extent, this assumption is also appropriate for providing critical guidelines in real applications. For classification problems, many important findings can be revealed from a study on Gaussian distributions.

The following numerical examples are specifically designed for demonstrating the intrinsic differences between Bayesian and mutual-information classifiers on Gaussian distributions. For calculations of NI's values on the following example, an open-source toolkit [39] is adopted for computations of mutual-information classifiers.

Example 1: Two cross-over points. The data for no rejection are given below:

$$\begin{aligned}&No\ rejection:\\ &\mu_1 = -1,\ \sigma_1 = 2,\ p(t_1) = 0.5,\ \lambda_{11} = 0,\ \lambda_{12} = 1,\\ &\mu_2 = 1,\quad \sigma_2 = 1,\ p(t_2) = 0.5,\ \lambda_{21} = 1,\ \lambda_{22} = 0\end{aligned}$$

The cost terms are used for Bayesian classifiers, but not for mutual-information classifiers. Table IV lists the results for both classifiers. One can obtain the same results when inputing $\lambda_{13} = 1 - \lambda_{23}$ for Bayesian classifiers. This is why a two-by-two matrix has to be used in the case of no rejection. Two cross-over points are formed in this examples (Fig. 1b). If no rejection is selected, both classifiers will have two boundary points. Bayesian classifiers will partition the classification regions by having $x_{b1} = x_{c1} = -0.238$ and $x_{b2} = x_{c2} = 3.57$. Mutual-information classifiers widen the region R_2 by $x_{b1} = -0.674$ and $x_{b2} = 4.007$ so that the error for Class 2 is much reduced. If considering zero costs for correct classifications and using eq. (18) with $\delta_i = \lambda_{21}/\lambda_{12}$, one can calculate a cost ratio below for an independent parameter to Bayesian classifiers in the case of no rejection:

$$\Lambda_{21} = \frac{\lambda_{21}}{\lambda_{12}} = \frac{p(x = x_b|t_1)p(t_1)}{p(x = x_b|t_2)p(t_2)}, \tag{50}$$

which is used to establish an equivalence between mutual-information classifiers and Bayesian classifiers. Substituting the boundary points of mutual-information classifier at $x_{b1} = -0.674$ and $x_{b2} = 4.007$ into $p(x|t_i)$ and (50), respectively, one receives a unique cost ratio value, $\Lambda_{21} = 2.002$. Hence, this mutual-information classifier has its unique equivalence to a specific Bayesian classifier which is exerted by the following conditions to the cost terms:

$$\lambda_{11} = 0,\ \lambda_{12} = 1.0,\ \lambda_{21} = 2.002,\ \lambda_{22} = 0.$$

Following the similar analysis above, one can reach a consistent observation for conducting a parametric study on σ_1/σ_2 in binary classifications. When two classes are well balanced, that is, $p(t_1) = p(t_2)$, both types of classifiers will produce larger errors in association with the larger-variance class. However, mutual-information classifiers always add more cost weight on the misclassification from a smaller-variance class. In other words, mutual-information classifiers prefer to generate a smaller error on a smaller-variance class in comparison with Bayesian classifiers when using zero-one cost functions (Table IV). This performance behavior seems closer to our intuitions in binary classifications under the condition of a balanced class dataset. When two classes are significantly different from their associated variances, a smaller-variance class generally represents an interested signal embedded within noise which often has a larger variance. The common practices in such classification scenarios require a larger cost weight on the misclassification from a smaller-variance class, and vice verse from a larger-variance class.

If a reject option is enforced for the following data:

$$\begin{aligned}&Rejection:\\ &\mu_1 = -1, \sigma_1 = 2, p(t_1) = 0.5, \lambda_{11} = 0, \lambda_{12} = 1.2,\\ &\lambda_{13} = 0.2,\\ &\mu_2 = 1, \sigma_2 = 1, p(t_2) = 0.5, \lambda_{21} = 1, \lambda_{22} = 0,\\ &\lambda_{23} = 0.6\end{aligned}$$

four boundary points are required to determine classification regions as shown in Fig 1d. For the given cost terms, a Bayesian classifier shows a lower error rate and a lower reject rate. While the rejects are almost equal between two classes, the errors are significantly different. One is able to adjust the errors and rejects by changing cost terms. For mutual-information classifiers, however, a balance is automatically made among error types and reject types. The results, shown in

TABLE IV
RESULTS OF EXAMPLE 1 ON UNIVARIATE GAUSSIAN DISTRIBUTIONS

Reject Option	Classifier Type	E_1 / E_2	E	Rej_1 / Rej_2	Rej	T_{r1} / T_{r2}	x_{b1}, x_{b2} / x_{b3}, x_{b4}	NI
No Rejection	Bayesian	0.170 / 0.057	**0.227**	0 / 0	0	- / -	-0.238, 3.571 / -, -	0.245
	Mutual-Information	0.215 / 0.024	0.239	0 / 0	0	- / -	-0.674, 4.007 / -, -	**0.260**
Rejection	Bayesian	0.131 / 0.024	**0.155**	0.083 / 0.084	0.167	0.333 / 0.375	-0.673, 0.162 / 3.171, 4.006	0.285
	Mutual-Information	0.154 / 0.006	0.160	0.118 / 0.068	0.186	0.141 / 0.445	-1.24, -0.0762 / 3.409, 4.571	**0.297**

Table IV, are considered for carrying the feature of objectivity in evaluations since no cost terms are specified subjectively. Note that a reject option enables both classifiers to reach higher values on their NI's than those in the case of without rejection. Because no "*one-to-one*" relations exist among the thresholds and the cost terms in a rejection case, one will fail to acquire a unique set of the equivalent cost terms between the Bayesian classifier and the mutual information classifier. For example, two sets of cost terms below will produce the same Bayesian classifiers based on the given solutions of the mutual information classifier:

$$\begin{cases} \lambda_{11}=0,\ \lambda_{12}=1,\ \lambda_{13}=0.0376, \\ \lambda_{21}=1,\ \lambda_{22}=0,\ \lambda_{23}=0.772 \end{cases}$$

or

$$\begin{cases} \lambda_{11}=0,\ \lambda_{12}=2.247,\ \lambda_{13}=1, \\ \lambda_{21}=7.069,\ \lambda_{22}=0,\ \lambda_{23}=1. \end{cases}$$

The meanings for two sets of cost terms are different. The first set indicates the same costs for errors, but the second one suggests the same costs for rejects. The results above imply an intrinsic problem of "*non-consistency*" for interpreting cost terms. One needs to be cautious about this problem when setting cost terms to Bayesian classifiers. This phenomenon occurs only in the case that a reject option is considered, but does not in the case without rejection. If the knowledge about thresholds exists, abstaining classifiers are better to apply T_{rk} directly for the input data (Table III), instead of employing cost terms. If no information is given about the thresholds or cost terms, mutual-information classifiers are able to provide an objective, or initial, reference of T_{rk} for Bayesian classifiers in cost sensitive learning.

Example 2: One cross-over point. The given inputs in this example are:

$$\begin{aligned} &No\ rejection: \\ &\mu_1=-1,\ \sigma_1=1,\ \lambda_{11}=0,\ \lambda_{12}=1, \\ &\mu_2=1,\ \sigma_2=1,\ \lambda_{21}=1,\ \lambda_{22}=0, \\ &p(t_1)=0.5, 2/3, 0.8, 0.9, 0.99, 0.999, 0.9999 \\ &p(t_2)=0.5, 1/3, 0.2, 0.1, 0.01, 0.001, 0.0001 \end{aligned}$$

Specific attention is paid to the class imbalanced data. When Class 2 alters from "*balanced*", "*minority*" to "*rare*" status in the whole data, we need to find out what behaviors both types of classifiers will display. For this purpose, a natural scheme with zero-one cost terms is set for Bayesian classifiers. Numerical investigations are conducted in this example. Table V lists the results of classifiers on the given data. If following the conventional term FNR for "*false negative rate*" in binary classifications, which is defined as:

$$FNR=\frac{E_2}{p(t_2)} \tag{51}$$

one can examine behaviors of FNR with respect to the ratio $p(t_1)/p(t_2)$. Sometimes, FNR is also called a "*miss rate*" [4]. Two types of classifiers show the same results when two classes are exactly balanced, that is, $p(t_1)/p(t_2)=1$. A single boundary point (Fig. 1a) separates two classes at the exact cross-over point ($x_b=x_c=0$). When one class, say $p(t_2)$ for Class 2, becomes smaller, the boundary point of Bayesian classifier moves toward to the mean point ($\mu_2=1$) of Class 2 (as pointed out in [[4], page 39]), and passes it finally. For keeping the smallest error, a Bayesian classifier will sacrifice the minority class. The results in Table V confirm Theorem 3 numerically on the Bayesian classifiers. Fig. 3 shows such behavior from the plot of "$E_2/p(t_2)$ vs. $p(t_1)/p(t_2)$". Note that the plots for the range from 10^{-4} to 10^0 on the $p(t_1)/p(t_2)$ axis are also depicted based on the data in Table V. For example, at the data point of $p(t_1)/p(t_2)=1/2$, one can get $E_2/p(t_2)=0.0594/(2/3)$, where 0.0594 is taken from E_1 for the data at $p(t_1)/p(t_2)=2$. The response of $E_2/p(t_2)$, representing the false negative rate, shows a distinguished property of Bayesian classifiers. One can observe that the complete set of Class 2 could be misclassified when it becomes extremely rare. This finding explains another reason for the question: "*Why do classifiers perform worse on the minority class?*" in [15].

Mutual-information classifiers exhibit different behavior in the given dataset. The first important feature is that the boundary point will shift toward the mean point ($\mu_2=1$) of Class 2 but will never go over it. The second feature informs that the response of $E_2/p(t_2)$ approaches asymptotically to a stable value, about 0.345 in this example, for a large ratio of $p(t_1)/p(t_2)$. This feature indicates that mutual-information classifiers will never sacrifice a minority class completely in this specific example. A significant fraction of the rare class is identified correctly. Moreover, the curve of $E_2/p(t_2)$ also demonstrates a lower, yet non-zero, bound on error rate (about 0.054) when $p(t_1)/p(t_2)$ approaches to zero. This phenomenon implies that, for Gaussian distributions of

TABLE V
RESULTS OF EXAMPLE 2 ON UNIVARIATE GAUSSIAN DISTRIBUTIONS

Classifier Type	$p(t_1)/p(t_2)$ [$p(t_1), p(t_2)$]	1 [$0.5, 0.5$]	2 [$2/3, 1/3$]	4 [$0.8, 0.2$]	9 [$0.9, 0.1$]	99 [$0.99, 0.01$]	999 [$0.999, 0.001$]	9999 [$0.9999, 0.0001$]
Bayesian	E_1	0.0793	0.0594	0.0362	0.0161	0.483e-3	0.422e-5	0.000
	E_2	0.0793	0.0856	0.0759	0.0539	0.903e-2	0.993e-3	0.1e-3
	$E_2/p(t_2)$	0.159	0.257	0.379	0.539	0.903	0.993	1.000
	$x_b(= x_c)$	0.0	0.347	0.693	1.10	2.30	3.45	4.61
	$H(T\|Y)$	0.631	0.591	0.491	0.349	0.0756	0.0113	0.00147
	NI	0.369	0.356	0.320	0.256	0.0644	0.00524	0.124e-3
Mutual-Information	E_1	0.0793	0.0867	0.0852	0.0772	0.0585	0.0551	0.0547
	E_2	0.0793	0.0637	0.0451	0.0264	0.331e-2	0.343e-3	0.345e-4
	$E_2/p(t_2)$	0.159	0.191	0.225	0.264	0.331	0.343	0.345
	x_b	0.0	0.126	0.246	0.367	0.562	0.597	0.601
	$H(T\|Y)$	0.631	0.586	0.472	0.320	0.0629	0.00957	0.00129
	NI	0.369	0.362	0.346	0.317	0.222	0.161	0.125

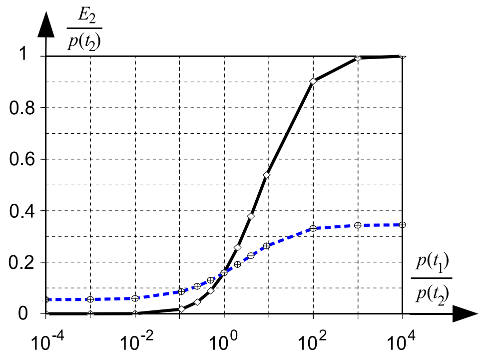


Fig. 3. Curves of "$E_2/p(t_2)$ vs. $p(t_1)/p(t_2)$". Solid curve: Bayesian classifier. Dashed curve: Mutual-information classifier.

classes, mutual-information classifiers generally do not hold a tendency of sacrificing a complete class in classifications. However, from a theoretical viewpoint, we still need to establish an analytical derivation of lower and upper bounds of $E_i/p(t_i)$ for mutual-information classifiers.

Example 3: Zero cross-over points. The given data for two classes are:

$$\mu_1 = 0,\ \sigma_1 = 2,\ p(t_1) = 0.8,$$
$$\mu_2 = 0,\ \sigma_2 = 1,\ p(t_2) = 0.2.$$

Although no data are specified to the cost terms, it generally implies a zero-one lost function for them [4]. From eq. (18), one can see a case of zero cross-over point occurs in this example (Fig. 4c). For the zero-one setting to cost terms, the Bayesian classifier will produce a specific classification result of "*Majority-taking-all*", that is, for all patterns identified as Class 1. The error gives to Class 2 only, and it holds the relation of $NI = 0$, which indicates that no information is obtained from the classifier [9]. One can imagine that the given example may describe a classification problem where a target class, with Gaussian distribution, is also corrupted with wider-band Gaussian noise in a frequency domain (Fig. 4a). The plots of $p(t_i)p(x|t_i)$ shows the overwhelming distribution of Class 1 over that of Class 2 (Fig. 4b). The plots on the posterior probability $p(t_i|x)$ indicate that Class 2 has no chance to be considered in the complete domain of x (Fig. 4c).

Table VI lists the results for both types of classifiers. The Bayesian approach fails to achieve the meaningful results on the given data. When missing input data of λ_{13} and λ_{23}, one cannot carry out the Bayesian approach for abstaining classifications. On the contrary, without specifying any cost term, mutual-information classifiers are able to detect the target class with a reasonable degree of accuracy. When no rejection is selected, less than two percentage error $(E2 = 1.53\%)$ happens to the target class. Although the total error $(E = 51.4\%)$ is much higher than its Bayesian counterpart $(E = 20\%, FNR = 0\%)$, the result of about eight percentage point $(FNR = 7.65\%)$ of the miss rate to the target is really meaningful in applications. If a reject option is engaged, the miss rate is further reduced to $FNR = 4.10\%$, but includes adding a reject rate of 29.1% over total possible patterns. This example confirms again the unique feature of mutual-information classifiers. The results of T_{rk} from mutual-information classifiers can also serve a useful reference for the design of Chow's abstaining classifiers, either with or without knowledge about cost terms.

C. Comparisons on Univariate Uniform Distributions

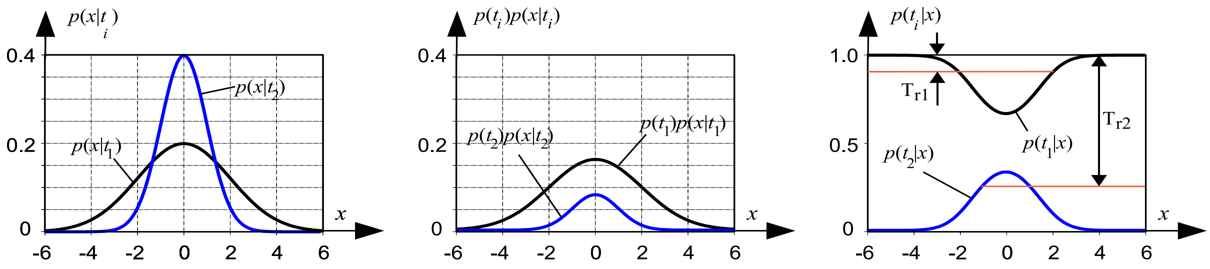
Uniform distributions are very rare in classification problems. This section shows one example given from [3]. A specific effort is made on numerical comparisons between the two types of classifiers.

Example 4: Partially overlapping between two distributions. The task for this example is to set the cost terms for controlling the decision results on the overlapping region for the given data from [3]:

$$p(x|t_1) = \begin{cases} 1 & when\ 0 \le x \le 1 \\ 0 & otherwise \end{cases}$$

$$p(x|t_2) = \begin{cases} 1/2 & when\ 0.5 \le x \le 2.5 \\ 0 & otherwise \end{cases}$$

$$p(t_1) = p(t_2) = 0.5.$$



(a) Plots of $p(x|t_i)$. (b) Plots of $p(t_i)p(x|t_i)$. (c) Plots of $p(t_i|x)$.

Fig. 4. Plots for Example 3 where (b)-(c) describe a signal (blue curve) embedded by wider-band noise (black curve).

TABLE VI
RESULTS OF EXAMPLE 3 ON UNIVARIATE GAUSSIAN DISTRIBUTIONS

Reject Option	Classifier Type	E_1 E_2	E	Rej_1 Rej_2	Rej	T_{r1} T_{r2}	x_{b1}, x_{b2} x_{b3}, x_{b4}	NI
No Rejection	Bayesian	0.0 0.2	0.2	0 0	0	- -	-, - -, -	0.0
	Mutual Information	0.499 0.0153	0.514	0 0	0	- -	-1.77, 1.77 -, -	0.0803
Rejection	Mutual Information	0.316 0.00819	0.324	0.239 0.0520	0.291	0.0945 0.749	-2.04, -1.03 1.03, 2.04	**0.0926**

In uniform distributions, a single independent parameter will be sufficient for classifications. Table VII lists the different results with respect to T_r. Note that the present results have extended Chow's abstaining classifiers by adding one more decision case of $f(x \in R_i) = y_2$ than those in [3]. The extension is attributed to the three rules used in eq. (10), rather than two in Chow's classifiers, which demonstrates a more general solution for classifications. One can see that mutual-information classifiers will decide $f(x \in R_i) = y_3$ from the given data of class distributions sine they receive the maximum value of NI. If no rejection is enforced, mutual-information classifiers will choose $f(x \in R_i) = y_1$ for their solution.

V. CONCLUSIONS

This work explored differences between Bayesian classifiers and mutual-information classifiers. Based on Chow's pioneering work [2][3], the author revisited Bayesian classifiers on two general scenarios for the reason of their increasing popularity in classifications. The first was on the zero-one cost functions for classifications without rejection. The second was on the cost distinctions among error types and reject types for abstaining classifications. In addition, the paper focused on the analytical study of mutual-information classifiers in comparison with Bayesian classifiers, which showed a basis for novel design or analysis of classifiers based on the entropy principle. The general decision rules were derived for both Bayesian and mutual-information classifiers based on the given assumptions. Two specific theorems were derived for revealing the intrinsic problems of Bayesian classifiers in applications under the two scenarios. One theorem described that Bayesian classifiers have a tendency of overlooking the misclassification error which is associated with a minority class. This tendency will degenerate a binary classification into a single class problem for the meaningless solutions. The other theorem discovered the parameter redundancy of cost terms in abstaining classifications. This weakness is not only on reaching an inconsistent interpretation to cost terms. The pivotal difficulty will be on holding the objectivity of cost terms. In real applications, information about cost terms is rarely available. This is particularly true for reject types. While Berger explained the demands for "*objective Bayesian analysis*" [43], we need to recognize that this goal may fail from applying cost terms in classifications. In comparison, mutual-information classifiers do not suffer such difficulties. Their advantages without requiring cost terms will enable the current classifiers to process abstaining classifications, like a new folder of "*Suspected Mail*" in Spam filtering [44]. Several numerical examples in this work supported the unique benefits of using mutual-information classifiers in special cases. The comparative study in this work was not meant to replace Bayesian classifiers by mutual-information classifiers. Bayesian and mutual-information classifiers can form "*complementary rather than competitive* (words from Zadeh [45])" solutions to classification problems. However, this work was intended to highlight their differences from theoretical studies. More detailed discussions to the differences between the two types of classifiers were given in Section IV. As a final conclusion, a simple answer to the question title is summarized below:

"*Bayesian and mutual-information classifiers are different essentially from their applied learning targets. From application viewpoints, Bayesian classifiers are more suitable to the cases when cost terms are exactly known for trade-off of error types and reject types. Mutual-information classifiers are capable of objectively balancing error types and reject types automatically without employing cost terms, even in the cases*

TABLE VII
RESULTS OF EXAMPLE 4 ON UNIVARIATE UNIFORM DISTRIBUTIONS

T_r	Decision on R_i	E_1, E_2	E	Rej_1, Rej_2	Rej	NI
$1/3 < T_r < 2/3$	$f(x \in R_i) = y_1$	0.0, 0.125	0.125	0, 0	0	0.549
$2/3 \leq T_r \leq 1$	$f(x \in R_i) = y_2$	0.250, 0	0.250	0, 0	0	0.311
$0 \leq T_r \leq 1/3$	$f(x \in R_i) = y_3$	0, 0	0	0.250, 0.125	0.375	**0.656**

of extremely class-imbalanced datasets, which may describe a theoretical interpretation why humans are more concerned about the accuracy of rare classes in classifications".

APPENDIX A
PROOF OF THEOREM 1

Proof: The decision rule of Bayesian classifiers for the "*no rejection*" case is well known in [4]. Then, only the rule for the "*rejection*" case is studied in the present proof. Considering eq. (6a) first from (5a), a pattern **x** is decided by a Bayesian classifier to be y_1 if $risk(y_1|\mathbf{x}) < risk(y_2|\mathbf{x})$ and $risk(y_1|\mathbf{x}) < risk(y_3|\mathbf{x})$. Substituting eqs. (1) and (2) into these inequality equations will result to:

$$\begin{aligned} Decide\ y_1\ if\ & \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} > \frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}} \\ and\ & \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} > \frac{\lambda_{21}-\lambda_{23}}{\lambda_{13}-\lambda_{11}}. \end{aligned} \tag{A1}$$

Similarly, one can obtain

$$\begin{aligned} Decide\ y_2\ if\ & \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} \leq \frac{\lambda_{21}-\lambda_{22}}{\lambda_{12}-\lambda_{11}} \\ and\ & \frac{p(\mathbf{x}|t_1)p(t_1)}{p(\mathbf{x}|t_2)p(t_2)} \leq \frac{\lambda_{23}-\lambda_{22}}{\lambda_{12}-\lambda_{13}}, \end{aligned} \tag{A2}$$

and eq. (6c) respectively. Eq. (A1) describes that a single upper bound within two boundaries will control a pattern **x** to be y_1. Similarly, eq. (A2) describes a lower bound for a pattern **x** to be y_2. From the constraints (3), one cannot determine which boundaries will be upper bound or lower bound. However, one can determine them from the following two hints in classifications:

A. Eq. (6c) describes a single lower boundary and a single upper boundary for a pattern **x** to be y_3.
B. The upper bound in (A1) and the lower bound in (A2) should be coincident with one of the boundaries in (6c) respectively so that classification regions from R_1 to R_3 will cover a complete domain of the pattern x (see Fig. 1c-d).

The hints above suggest the novel constraints for λ_{ij} as shown in eq. (6d). Any violation of the constraints will introduce a new classification region R_4, which is not correct for the present classification background. The constraints of thresholds for rejection (6e) can be derived directly from (6c) and (6d). ■

APPENDIX B
TIGHTER BOUNDS BETWEEN CONDITIONAL ENTROPY AND BAYESIAN ERROR IN BINARY CLASSIFICATIONS

In the study of relations between mutual information (I) and Bayesian error (E), two important studies are reported on the lower bound (LB) by Fano [48] and the upper bound (UB) by Kovalevskij [49] in the forms of

$$LB:\ E \geq \frac{H(T)-I(T,Y)-H(E)}{log_2(m-1)} = \frac{H(T|Y)-H(E)}{log_2(m-1)}, \tag{B1}$$

$$UB:\ E \leq \frac{H(T)-I(T,Y)}{2} = \frac{H(T|Y)}{2}, \tag{B2}$$

where m is the total number of classes in T, $H(E)$ is the binary Shannon entropy, and $H(T|Y)$ is called conditional entropy which can be derived from a general relation [4]:

$$I(T,Y) = I(Y,T) = H(T) - H(T|Y) = H(Y) - H(Y|T). \tag{B3}$$

For binary classifications $(m = 2)$, a tighter Fano's bound in [50] [51] is adopted. Based on the rationals of Bayesian error, we suggest the tighter upper and lower bounds in the forms of:

$$Modified\ LB:\ H(E) \geq H(T|Y),\ and\ 0 \leq E, \tag{B4}$$

$$Modified\ UB:\ E \leq min(p(t_1), p(t_2), \frac{H(T|Y)}{2}). \tag{B5}$$

Fig. 5 shows the bounds in binary classifications, which is different from "$I(T,Y)\ vs.\ E$" plots in [51]. Because of the equivalent relations [11]:

$$max\ I(T,Y) = min\ H(T|Y), \tag{B6}$$

the plots for $H(T|Y)$ is preferable, which does not require the information of $H(T)$. One is able to draw the lower-bound curve from (B4), but unable to show its explicit form for E. The areal feature of the enclosed bounds suggests two important properties about the relations. The first is due to the approximations in the derivations of the bounds [48] [49]. The second represents an intrinsic property of no "*one-to-one*" relations between mutual information and accuracy in classifications [10].

Triangles and circles shown in Fig. 5 represent the paired data in Table V from Bayesian classifiers and mutual information classifiers, respectively. They clearly demonstrate the specific forms in their positions within the same pairs. The circle position is either coincident or "*up and/or left*" to its counterpart. These forms are attributed to the different directions of driving force for two types of classifiers. One is for "$min\ E$" and the other for "$min\ H(T|Y)$".

Important findings are observed in related to the bounds. First, the triangles demonstrate Fano's bound in eq. (B4) to be a very tight lower bound. Second, an upper bound of E_{max} exists according to Theorem 3, which is tighter than a constant one ($= 0.5$) in [50]. When p_{min} decreases as shown in Table V, the upper bound from the maximum Bayesian error

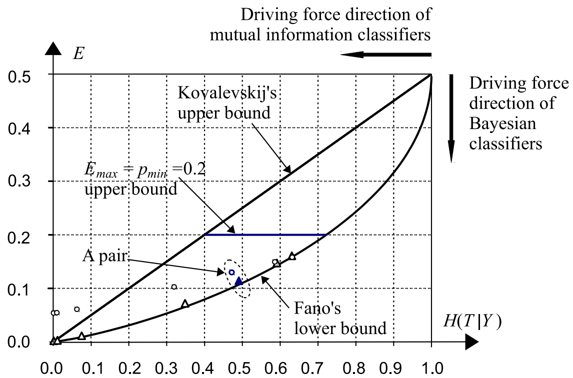


Fig. 5. The bounds between conditional entropy $H(T|Y)$ and Bayesian error in binary classifications. Triangles and circles are the data in Table V from Bayesian classifiers and mutual information classifiers, respectively. An upper bound from the maximum Bayesian error exists, say, $E_{max} = 0.2$ for the filled triangle.

will become closer to its associated data. Third, the Fano's lower bound is effective for all classifiers, including mutual information classifiers. However, the upper bounds, even the constant one ($= 0.5$) becomes invalid for mutual information classifiers (see the data $E = 0.514$ in Table VI).

The observations above indicate the necessity of further investigation into the upper bounds for better descriptions of the relations. If much tighter upper bounds are possible, they are desirable to disclose their theoretical insights between the two types of classifiers.

ACKNOWLEDGMENTS

This work was supported in part by NSFC #61075051. The assistances from Mr. Yajun Qu and Mr. Christian Ocier in the text preparation are gratefully acknowledged.

REFERENCES

[1] S.R. Kulkarni, G. Lugosi, G., and S.S. Venkatesh, "Learning Pattern Classification-A Survey", *IEEE Transactions on Information Theory*, vol. 44, pp. 2178 - 2206, 1998.
[2] C. Chow, "An Optimum Character Recognition System Using Decision Functions," *IRE Trans. Electronic Computers*, vol. 6, pp. 247-254, 1957.
[3] C. Chow, "On Optimum Recognition Error and Reject Tradeoff," *IEEE Trans. Information Theory*, vol. 16, pp. 41-46, 1970.
[4] R.O. Duda, P.E. Hart, and D. Stork, *Pattern Classification*, 2nd eds., John Wiley: New York, 2001.
[5] J.R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81-106, 1986.
[6] T.O. Kvålseth, "Entropy and Correlation: Some Comments," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 17, pp. 517-519, 1987.
[7] T.D. Wickens, *Multiway Contingency Tables Analysis for the Social Sciences*, Lawrence Erlbaum, Hillsdale, New Jersey, 1989.
[8] J.C. Principe, J.W. Fisher III, and D. Xu, "Information Theoretic Learning," *Unsupervised Adaptive Filtering*, Wiley: New York, pp. 265-319, 2000.
[9] D.J.C. Mackay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press: Cambridge, 2003.
[10] Y. Wang and B.-G. Hu, "Derivations of Normalized Mutual Information in Binary Classifications," *Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 155-163, 2009.
[11] B.-G. Hu and Y. Wang, "Evaluation Criteria Based on Mutual Information for Classifications Including Rejected Class," *Acta Automatica Sinica*, vol. 34, pp. 1396-1403, 2008.
[12] P. Domingos, "MetaCost: A General Method for Making Classifiers Cost-sensitive," *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155-164, 1999.
[13] C. Elkan, "The Foundations of Cost-sensitive Learning," *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, pp. 973-978, 2001.
[14] N. Japkowicz and S. Stephen, "The Class Imbalance Problem: A Systematic Study", *Intelligent Data Analysis*, vol. 6, pp. 429-449, 2002.
[15] G.M. Weiss and F. Provost, "The Effect of Class Distribution on Classifier Learning," Technical Report ML-TR 43, Department of Computer Science, Rutgers University, 2001.
[16] N.V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial to the Special Issue on Learning from Imbalanced Data Sets," *ACM SIGKDD Explorations*, vol. 6, pp. 1-6, 2004.
[17] Z.-H, Zhou and X.-Y. Liu, "Training Cost-sensitive Neural Networks with Methods Addressing the Class Imbalance Problem", *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, pp. 63-77, 2006.
[18] H. He and E.A. Garcia, "Learning from Imbalanced Data", *IEEE Transactions on Knowledge and Data Engineering*, vol. 21,pp. 1263-1284, 2009.
[19] Q. Yang and X. Wu, "10 Challenging Problems in Data Mining Research", *International Journal of Information Technology and Decision Making*, vol. 5, pp. 597-604, 2006.
[20] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Chapman and Hall/CRC Press: Boca Raton, FL, 1984.
[21] B. Dubuisson and M. Masson, "A Statistical Decision Rule with Incomplete Knowledge about Classes," *Pattern Recognition*, vol. 26, pp. 155-165, 1993.
[22] T.M. Ha, "The Optimum Class-Selective Rejection Rule", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 608-615, 1997.
[23] M. Golfarelli, D. Maio, and D. Malton, "On the Error-reject Trade-off in Biometric Verification Systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 786-796, 1997.
[24] Y. Baram, "Partial Classification: The Benefit of Deferred Decision," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, pp. 769-776, 1998.
[25] C. Ferri, and J. Hernandez-Orallo, "Cautious Classifiers," *Proceedings of the 1st International Workshop on ROC Analysis in Artificial Intelligence (ROCAI-2004)*, pp. 27-36, 2004.
[26] F. Tortorella, "A ROC-based Reject Rule for Dichotomizers," *Pattern Recognition Letters*, vol. 25, pp. 167-180, 2005.
[27] C.C. Friedel, U. Ruckert, and S. Kramer, "Cost Curves for Abstaining Classifiers," *Proceedings of the ICML-2006Workshop on ROC Analysis (ROCML-2006)*, pp. 33-40, 2006.
[28] T. Pietraszek, "On the Use of ROC Analysis for the Optimization of Abstaining Classifiers," *Machine Learning*, vol. 69, pp. 137-169, 2007.
[29] S. Vanderlooy, I.G. Sprinkhuizen-Kuyper, E.N. Smirnov and J. van den Herik, "The ROC Isometrics Approach to Construct Reliable Classifiers," *Intelligent Data Analysis*, vol. 13, pp. 3-37, 2009.
[30] E. Grall-Maës and P. Beauseroy, "Optimal Decision Rule with Class-selective Rejection and Performance Constraints," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, pp. 2073-2082, 2009.
[31] C.M. Santos-Perira and A.M. Pires, "On Optimal Reject Rules and ROC Curves," *Pattern Precognition Letters*, vol. 26, pp. 943-952, 2005.
[32] M. Yuan and M. Wegkamp, "Classification Methods with Reject Option Based on Convex Risk Minimization", *Journal of Machine Learning Research*, vol. 11, pp. 111 - 130, 2010.
[33] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press: New York, 1990.
[34] G. Fumera, F. Roli, and G. Giacinto, "Reject Option with Multiple Thresholds," *Pattern Recognition*, vol. 33, pp. 2099-2101, 2000.
[35] S.H. Yang, B.-G. Hu, and P.-H. Counede, "Structural Identifiability of Generalized Constraints Neural Network Models for Nonlinear Regression," *Neurocomputing*, vol. 72, pp. 392-400, 2008.
[36] B.-G. Hu, H.-B. Qu, Y. Wang, and S.-H. Yang, "A Generalized Constraint Neural Networks Model: Associating Partially Known Relationships for Nonlinear Regressions," *Information Sciences*, vol. 179, pp. 1929-1943, 2009.
[37] C.E. Shannon, "A Mathematical Theory of Communication", *Bell System Technical Journal*, vol. 27, pp. 379-423 and pp. 623-656, 1948.
[38] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, 2nd ed., John Wiley: New York, 2006.
[39] B.-G. Hu, "Information Measure Toolbox for Classifier Evaluation on Open Source Software Scilab," *Proceedings of IEEE International Workshop on Open-source Software for Scientific Computation (OSSC-2009)*, pp. 179-184, 2009.

[40] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer-Verlag: New York, 1985.

[41] I. Kononenko and I. Bratko, “Information-based Evaluation Criterion for Classifier’s Performance”, *Machine Learning*, vol. 6, pp. 67-80, 1991.

[42] B. Liu, “A Survey of Entropy of Fuzzy Variables,” *Journal of Uncertain Systems*, vol.1, pp.4-13, 2007.

[43] J.O. Berger, “The Case for Objective Bayesian Analysis”, *Bayesian Analysis*, vol. 1, pp. 385-402, 2006.

[44] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, “A Bayesian Approach to Filtering Junk E-mail”. AAAI’98 Workshop on Learning for Text Categorization, 1998.

[45] L.A. Zadeh, “Discussion: Probability Theory and Fuzzy Logic are Complementary rather than Competitive”, *Technometrics*, vol. 37, pp. 271-276, 1995.

[46] R. Santos-Rodríguez, A. Guerrero-Curieses, R. Alaiz-Rodríguez, and J. Cid-Sueiro, “Cost-Sensitive Learning Based on Bregman Divergences”, *Machine Learning*, vol. 76, pp. 271-285, 2009.

[47] Q. Fu, D. S. Mansjur, and B.-H. Juang, “Empirical System Learning for Statistical Pattern Recognition With Non-Uniform Error Criteria”, *IEEE Transactions on Signal Processing*, vol. 58, pp. 4621 - 4633, 2010.

[48] R.M. Fano, *Transmission of Information: A Statistical Theory of Communication*, MIT: New York, 1961.

[49] V.A. Kovalevskij, “The problem of character recognition from the point of view of mathematical statistics”, *Character Readers and Pattern Recognition*, pp. 3-30, Spartan: New York, 1968. Russian edition 1965.

[50] I. Vajda and J. Zvárová, “On generalized entropies, Bayesian decisions and statistical diversity”, *Kybernetika*, pp. 675-696, 2007.

[51] J.W. Fisher III, M. Siracusa, and K. Tieu, “Estimation of signal information content for classification”, *Proceedings of IEEE DSP Workshop*, pp. 353-358, 2009.